Lasso and probabilistic inequalities for multivariate point processes

Niels Richard Hansen, Patricia Reynaud-Bouret, Vincent Rivoirard

Submitted to the Bernoulli
arXiv: arXiv:1208.0570

Lasso and probabilistic inequalities for multivariate point processes

NIELS RICHARD HANSEN[1], PATRICIA REYNAUD-BOURET[2] and VINCENT RIVOIRARD[3]

[1]*Department of Mathematical Sciences, University of Copenhagen, Universitetsparken 5, 2100 Copenhagen, Denmark. E-mail:* Niels.R.Hansen@math.ku.dk

[2]*Univ. Nice Sophia Antipolis, CNRS, LJAD, UMR 7351, 06100 Nice, France. E-mail:* Patricia.Reynaud-Bouret@unice.fr

[3]*CEREMADE, CNRS-UMR 7534, Université Paris Dauphine, Place Maréchal de Lattre de Tassigny, 75775 Paris Cedex 16, France. INRIA Paris-Rocquencourt, projet Classic E-mail:* Vincent.Rivoirard@dauphine.fr

Due to its low computational cost, Lasso is an attractive regularization method for high-dimensional statistical settings. In this paper, we consider multivariate counting processes depending on an unknown function parameter to be estimated by linear combinations of a fixed dictionary. To select coefficients, we propose an adaptive ℓ_1-penalization methodology, where data-driven weights of the penalty are derived from new Bernstein type inequalities for martingales. Oracle inequalities are established under assumptions on the Gram matrix of the dictionary. Non-asymptotic probabilistic results for multivariate Hawkes processes are proven, which allows us to check these assumptions by considering general dictionaries based on histograms, Fourier or wavelet bases. Motivated by problems of neuronal activity inference, we finally carry out a simulation study for multivariate Hawkes processes and compare our methodology with the *adaptive Lasso procedure* proposed by Zou in [64]. We observe an excellent behavior of our procedure. We rely on theoretical aspects for the essential question of tuning our methodology. Unlike adaptive Lasso of [64], our tuning procedure is proven to be robust with respect to all the parameters of the problem, revealing its potential for concrete purposes, in particular in neuroscience.

Keywords: Multivariate counting process, Hawkes processes, adaptive estimation, Lasso procedure, Bernstein-type inequalities.

1. Introduction

The Lasso, proposed in [58], is a well established method that achieves sparsity of an estimated parameter vector via ℓ_1-penalization. In this paper, we focus on using Lasso to select and estimate coefficients in the basis expansion of intensity processes for multivariate point processes.

Recent examples of applications of multivariate point processes include the modeling of multivariate neuron spike data [43, 47], stochastic kinetic modeling [7] and the modeling of the distribution of ChIP-seq data along the genome [20]. In the previous examples, the intensity of a future occurrence of a point depends on the history of all or some of the coordinates of the point processes, and it is of particular interest to estimate this dependence. This can be achieved using a parametric family of models, as in several of the papers above. Our aim is to provide a nonparametric method based on the Lasso.

The statistical properties of Lasso are particularly well understood in the context of regression with i.i.d. errors or for density estimation where a range of *oracle inequalities* have been established. These inequalities, now widespread in the literature, provide theoretical error bounds that hold on events with a controllable (large) probability. See for instance [5, 6, 15, 16, 17, 18, 60]. We refer the reader to [13] for an excellent account on many state-of-the-art results. One main challenge in this context is to obtain as weak conditions as possible on the design – or Gram – matrix. The other important challenge is to be able to provide an ℓ_1-penalization procedure that provides excellent performance from both theoretical and practical points of view. Standard Lasso proposed in [58] and based on deterministic constant weights constitutes a major contribution from the methodological point of view, but underestimation due to its shrinkage nature may lead to poor practical performance in some contexts. Alternative two step procedures have been suggested to overcome this drawback (see [42, 61, 64]). Zou in [64] also discusses problems for standard Lasso to cope with variable selection and consistency simultaneously. He overcomes these problems by introducing non-constant data-driven ℓ_1-weights based on preliminary consistent estimates.

1.1. Our contributions

In this paper we consider an ℓ_1-penalized least squares criterion for the estimation of coefficients in the expansion of a function parameter. As in [5, 35, 61, 64], we consider non-constant data-driven weights. However the setup is here that of multivariate point processes and the function parameter that lives in a Hilbert space determines the point process intensities. Even in this unusual context, the least squares criterion also involves a random Gram matrix as well, and in this respect, we present a fairly standard oracle inequality with a strong condition on this Gram matrix (see Theorem 1 in Section 2).

One major contribution of this article is to provide probabilistic results that enable us to calibrate ℓ_1-weights in the most general setup (see Theorem 2 in Section 2). This is naturally linked to sharp Bernstein type inequalities for martingales. In the literature, those kinds of inequalities generally provide upper bounds for the martingale that are deterministic and unobservable [57, 59]. To choose data-driven weights we need observable bounds. More recently, there have been some attempts to use self-normalized processes in order to provide more flexible and random upper bounds [4, 25, 26, 27]. Nevertheless, those bounds are usually not (completely) observable when dealing with counting processes. We prove a result that goes further in this direction by providing a completely sharp random observable upper bound for the martingale in our counting process framework (see Theorem 3 in Section 3).

The second main contribution is to provide a quite theoretical and abstract framework to deal with processes whose intensity is (or is well approximated by) a linear transformation of deterministic parameters to infer. This general framework also allows for different asymptotics in terms of the number of observed processes or in terms of the duration of the recording of observations, according to the setup. We focus in this paper on three main examples: the Poisson model, the Aalen multiplicative intensity model and the multivariate Hawkes process, but many other situations can be expressed in the present framework, which has the advantage of full flexibility. The first two examples have been extensively studied in the literature as we detail hereafter, but Hawkes processes are typical of situations where very little is known from a nonparametric point of view, and where fully implementable adaptive methods do not exist until the present work, to the best of our knowledge. They also constitute processes that are often used in practice – in particular in neuroscience – as explained below.

It is also notable that we, in each of these three previous examples, can verify explicitly if the strong condition on the Gram matrix mentioned previously is fulfilled with probability close to 1 (see Section 4 for the Poisson and Aalen cases and Section 5 for the Hawkes case). For the multivariate Hawkes process this involves novel probabilistic inequalities. Even though the Hawkes processes have been studied extensively in the literature, see [9, 24], very little is known about exponential inequalities and non-asymptotic tail control. Besides the univariate case [51], no exponential inequality controlling the number of points per interval is known to us. We derive such results and other sharp controls on the convergence in the ergodic theorem to obtain control on the Gram matrix.

Finally, we carry out a simulation study in Section 6 for the most intricate process, namely the multivariate Hawkes process, with a main aim: to convince practitioners, for instance in neuroscience, that this method is indeed fully implementable and gives good results in practice. Data-driven weights for practical purposes are slight modifications of theoretical ones. These modifications essentially aim at reducing the number of tuning parameters to one. Due to non-negligible shrinkage that is unavoidable, in particular for large coefficients, we propose a two step procedure where estimation of coefficients is handled by using ordinary least squares estimation on the support preliminary determined by our Lasso methodology. Tuning issues are extensively investigated in our simulation study, and Table 1 in Section 6.3 shows that our methodology can easily and robustly be tuned by using limit values imposed by assumptions of Theorem 2. We naturally compare our procedure with *adaptive Lasso* of [64] for which weights are proportional to the inverse of ordinary least squares estimates. The latter is very competitive for estimation aspects since shrinkage becomes negligible if the preliminary OLS estimates are large. But adaptive Lasso does not incorporate random fluctuations of coefficient estimators. So it is most of the time outperformed by our procedure. In particular, tuning adaptive Lasso in the Hawkes setting is a difficult task, which cannot be tackled by using standard cross-validation. Our simulation study shows that the performance of adaptive Lasso is very sensitive to the choice of the tuning parameter. Robustness with respect to tuning is another advantage of our method over adaptive Lasso. For simulations, the framework of neuronal networks is used. Our short study proves that our methodology can be used for solving concrete problems in neuroscience such as the inference of functional connectivity graphs.

1.2. Multivariate counting process

The framework introduced here and used throughout the paper aims at unifying several situations, making the reading easier. Main examples are then shortly described, illustrating the use of this setup.

We consider an M-dimensional counting process $(N_t^{(m)})_{m=1,\dots,M}$, which can also be seen as a random point measure on $\mathbb{R}$ with marks in $\{1,\dots,M\}$, and corresponding *predictable* intensity processes $(\lambda_t^{(m)})_{m=1,\dots,M}$ under a probability measure $\mathbb{P}$ (see [8] or [24] for precise definitions).

Classical models assume that the intensity $\lambda_t^{(m)}$ can be written as a *linear predictable transformation* of a deterministic function parameter f_0 belonging to a Hilbert space $\mathcal{H}$ (the structure of $\mathcal{H}$, and then of f_0, will differ according to the context, as illustrated below). We denote this linear transformation by

$$\psi(f) = (\psi^{(1)}(f), ..., \psi^{(M)}(f)). \tag{1.1}$$

Therefore, for classical models, for any t,

$$\lambda_t^{(m)} = \psi_t^{(m)}(f_0). \tag{1.2}$$

The main goal in classical settings is to estimate f_0 based on observing $(N_t^{(m)})_{m=1,\dots,M}$ for $t \in [0,T]$. Actually, we do not require in Theorems 1 and 2 that (1.2) holds. Our aim is mainly to furnish an estimate of the best linear approximation $\psi_t^{(m)}(f_0)$ of the underlying intensity $\lambda_t^{(m)}$.

Let us illustrate the general setup with three main examples: First, the case with i.i.d. observations of an inhomogeneous Poisson process on $[0,1]$ and unknown intensity, second, the well known Aalen multiplicative intensity model and third, the central example of the multivariate Hawkes process. For the first two models, asymptotics is with respect to M (T is fixed). For the third one, M is fixed and asymptotics is with respect to T.

1.2.1. The Poisson model

Let us start with a very simple example which will be somehow a toy problem here compared to the other examples. In this example we take $T = 1$ and assume that we observe M i.i.d. Poisson processes on $[0,1]$ with common intensity $f_0 : [0,1] \longmapsto \mathbb{R}_+$. Asymptotic properties are obtained when M tends to infinity. In this case, the intensity $\lambda^{(m)}$ of the m'th process $N^{(m)}$ is f_0, which does not depend on m: Therefore for any $m \in \{1,\dots,M\}$ and any t, we set

$$\psi_t^{(m)}(f_0) := f_0(t),$$

and $\mathcal{H} = \mathbb{L}_2([0,1])$ is equipped with the classical norm defined by

$$\|f\| = \left(\int_0^1 f^2(t)\mathrm{d}t\right)^{1/2}.$$

This framework has already been extensively studied from an adaptive point of view: see for instance [48, 63] for model selection methods, [50] for wavelet thresholding estimation or [53] for kernel estimates. In this context, our present general result matches with existing minimax adaptation results where asymptotics is with respect to M.

1.2.2. The Aalen multiplicative intensity model

This is one of the most popular counting processes because of its adaptivity to various situations (from Markov models to censored lifetime models) and its various applications to biomedical data (see [2]). Given $\mathcal{X}$ a Hilbert space, we consider $f_0 : [0, T] \times \mathcal{X} \longmapsto \mathbb{R}_+$, and we set for any $t \in \mathbb{R}$,

$$\lambda_t^{(m)} = \psi_t^{(m)}(f_0) := f_0(t, X^{(m)}) Y_t^{(m)},$$

where $Y^{(m)}$ is an observable predictable process and $X^{(m)}$ represents covariates. In this case, $\mathcal{H} = \mathbb{L}_2([0, T] \times \mathcal{X})$. Our goal is to estimate f_0 and not to select covariates. So, to fix ideas one sets $\mathcal{X} = [0, 1]$ and $T = 1$. Hence $\mathcal{H}$ can be identified with $\mathbb{L}_2([0, 1]^2)$. For right-censored data, f_0 usually represents the hazard rate. The presence of covariates in this pure nonparametric model is the classical generalization of the semi-parametric model proposed by Cox (see [39] for instance). Note that the Poisson model is a special case of the Aalen model.

The classical framework consists in assuming that $(X^{(m)}, Y^{(m)}, N^{(m)})_{m=1,\dots,M}$ is an i.i.d. M-sample and as for the Poisson model, it is natural to investigate asymptotic properties when $M \to +\infty$. If there are no covariates, several adaptive approaches already exist: See [11, 12, 49] for various penalized least-squares contrasts and [21] for kernel methods in special cases of censoring. In the presence of covariates, one can mention [1, 2] for a parametric approach, [23, 39] for a model selection approach and [29] for a Lasso approach. Let us also cite [14] where covariate selection via penalized MLE has been studied. Once again, our present general result matches with existing oracle results. In [21], exponential control of random fluctuations leading to adaptive results are derived without using the martingale theory. In more general frameworks (as in [23] for instance), martingales are required and this even when i.i.d. processes are involved.

1.2.3. Hawkes processes

Hawkes processes are the point processes equivalent to autoregressive models. In seismology, Hawkes processes can model earthquakes and their aftershocks [62]. More recently they have been used to model favored or avoided distances between occurrences of motifs [32] or Transcription Regulatory Elements [20] on the DNA. We can also mention the use of Hawkes processes as models of social interactions [44] or financial phenomena [3].

In the univariate setting, with $M = 1$, the intensity of a nonlinear Hawkes process $(N_t)_{t>0}$ is given by

$$\lambda_t = \phi\left(\int_{-\infty}^{t-} h(t-u)dN_u\right),$$

where $\phi : \mathbb{R} \mapsto \mathbb{R}_+$ and $h : \mathbb{R}_+ \mapsto \mathbb{R}$ (see [9]). A particular case is Hawkes's self exciting point process, for which h is nonnegative and $\phi(x) = \nu + x$ where $\nu > 0$ (see [9, 24, 34]). For instance, for seismological purposes, ν represents the spontaneous occurrences of real original earthquakes. The function h models self-interaction: after a shock at time u, we observe an aftershock at time t with large probability if $h(t-u)$ is large.

These notions can be easily extended to the multivariate setting and in this case the intensity of $N^{(m)}$ takes the form:

$$\lambda_t^{(m)} = \phi^{(m)}\left(\sum_{\ell=1}^{M}\int_{-\infty}^{t-} h_\ell^{(m)}(t-u)\mathrm{d}N^{(\ell)}(u)\right).$$

Theorem 7 of [9] gives conditions on the functions $\phi^{(m)}$ (namely Lipschitz properties) and on the functions $h_\ell^{(m)}$ to obtain existence and uniqueness of a stationary version of the associated process. Throughout this paper, we assume that for any $m \in \{1,\dots,M\}$,

$$\phi^{(m)}(x) = (\nu^{(m)} + x)_+,$$

where $\nu^{(m)} > 0$ and $(\cdot)_+$ denotes the positive part. Note that in [20, 22], the case $\phi^{(m)}(x) = \exp(\nu^{(m)}+x)$ was studied. However, Lipschitz properties required in [9] are not satisfied in this case. By introducing, as previously, the linear predictable transformation $\psi(f) = (\psi^{(1)}(f), ..., \psi^{(M)}(f))$ with for any m and any t

$$\psi_t^{(m)}(f_0) := \nu^{(m)} + \sum_{\ell=1}^{M}\int_{-\infty}^{t-} h_\ell^{(m)}(t-u)\mathrm{d}N^{(\ell)}(u), \tag{1.3}$$

with $f_0 = (\nu^{(m)}, (h_\ell^{(m)})_{\ell=1,\dots,M})_{m=1,\dots,M}$, we have $\lambda_t^{(m)} = (\psi_t^{(m)}(f_0))_+$. Note that the upper integration limits in (1.3) are $t-$, that is, the integrations are all over the open interval $(-\infty, t)$. This assures predictability of the intensity disregarding the values of $h_\ell^{(m)}(0)$. Alternatively, it can be assumed throughout that $h_\ell^{(m)}(0) = 0$, in which case the integrals in (1.3) can be over $(-\infty, t]$ without compromising predictability. The parameter $\nu^{(m)}$ is called the *spontaneous rate*, whereas the function $h_\ell^{(m)}$ is called the *interaction function* of $N^{(\ell)}$ on $N^{(m)}$. The goal is to estimate f_0 by using Lasso estimates. In the sequel, we will assume that the support of $h_\ell^{(m)}$ is bounded. By rescaling we can then assume that the support is in $[0,1]$, and we will do so throughout. Note that in this case we will need to observe the process on $[-1, T]$ in order to compute $\psi_t^{(m)}(f_0)$ for $t \in [0, T]$. The Hilbert space $\mathcal{H}$ associated with this setup is

$$\mathcal{H} = (\mathbb{R} \times \mathbb{L}_2([0,1])^M)^M = \Big\{f = \Big((\mu^{(m)}, (g_\ell^{(m)})_{\ell=1,\dots,M})_{m=1,\dots,M}\Big) :$$
$$g_\ell^{(m)} \text{ with support in } [0,1] \text{ and } \|f\|^2 = \sum_m (\mu^{(m)})^2 + \sum_m\sum_\ell \int_0^1 g_\ell^{(m)}(t)^2\mathrm{d}t < \infty\Big\}.$$

Some theoretical results are established in this general setting but to go further, we shall consider in Section 5 the case where the functions $h_\ell^{(m)}$ are nonnegative and then $\lambda_t^{(m)}$ is a linear function of f_0, as in Sections 1.2.1 and 1.2.2:

$$\lambda_t^{(m)} = \psi_t^{(m)}(f_0).$$

The multivariate point process associated with this setup is called the *multivariate Hawkes self exciting point process* (see [34]). In this example, M is fixed and asymptotic properties are obtained when T tends to infinity.

To the best of our knowledge, nonparametric estimation for Hawkes models has only been proposed in [52] in the univariate setting where the method is based on ℓ_0-penalization of the least-squares contrast. However, due to ℓ_0-penalization, the criterion is not convex and the computational cost, in particular for the memory storage of all the potential estimators, is huge. Therefore, this method has never been adapted to the multivariate setting. Moreover, the penalty term in this method is not data-driven and ad-hoc tuning procedures have been used for simulations. This motivates the present work and the use of a convex Lasso criterion combined with data-driven weights, to provide a fully implementable and theoretically valid data-driven method, even in the multivariate case.

Applications in neuroscience. Hawkes processes can naturally be applied to model neuronal activity. Extracellular action potentials can be recorded by electrodes and the recorded data for the neuron m can be seen as a point process, each point corresponding to the peak of one action potential of this neuron (see [10] for instance for precise definitions). When M neurons are simultaneously recorded, one can assume that we are faced with a realization of $N = (N^{(m)})_{m=1,\dots,M}$ modeled by a multivariate Hawkes process. We then assume that the intensity associated with the activity of the neuron m is given by $\lambda_t^{(m)} = (\psi_t^{(m)}(f_0))_+$, where $\psi_t^{(m)}(f_0)$ is given in (1.3). At any occurrence $u < t$ of $N^{(\ell)}$, $\psi_t^{(m)}(f_0)$ increases (excitation) or decreases (inhibition) according to the sign of $h_\ell^{(m)}(t-u)$. Modeling inhibition is essential from the neurobiological point of view. So, we cannot assume that all interaction functions are nonnegative, and we cannot omit the positive part. More details are given in Section 6.

In neuroscience, Hawkes processes combined with maximum likelihood estimation have been used in the seminal paper [22], but the application of the method requires a too huge number of observations for realistic practical purposes. Models based on Hawkes processes have nevertheless been recently discussed in neuroscience, since they constitute one of the few simple models able to produce dependence graphs between neurons, that may be interpreted in neuroscience as functional connectivity graphs [45, 46]. However, many nonparametric statistical questions arise that are not solved yet in order to furnish a fully applicable tool for real data analysis [38]. We think that the Lasso-based methodology presented in this paper may furnish the first robust tool in this direction.

1.3. Notation and overview of the paper

Some notation from the general theory of stochastic integration is useful to simplify the otherwise quite heavy notation. If $H = (H^{(1)}, ..., H^{(M)})$ is a multivariate process with locally bounded coordinates, say, and $X = (X^{(1)}, ..., X^{(M)})$ is a multivariate semi-martingale, we define the real valued process $H \bullet X$ by

$$H \bullet X_t := \sum_{m=1}^{M} \int_0^t H_s^{(m)} \mathrm{d}X_s^{(m)}.$$

Given $\mathfrak{F} : \mathbb{R} \longmapsto \mathbb{R}$ we use $\mathfrak{F}(H)$ to denote the coordinatewise application of $\mathfrak{F}$, that is $\mathfrak{F}(H) = (\mathfrak{F}(H^{(1)}), ..., \mathfrak{F}(H^{(M)}))$. In particular,

$$\mathfrak{F}(H) \bullet X_t = \sum_{m=1}^{M} \int_0^t \mathfrak{F}(H_s^{(m)}) \mathrm{d}X_s^{(m)}.$$

We also define the following scalar product on the space of multivariate processes. For any multivariate processes $H = (H^{(1)}, ..., H^{(M)})$ and $K = (K^{(1)}, ..., K^{(M)})$, we set

$$< H, K >_{\mathrm{proc}} := \sum_{m=1}^{M} \int_0^T H_s^{(m)} K_s^{(m)} \mathrm{d}s,$$

the corresponding norm being denoted $\|H\|_{\mathrm{proc}}$. Since ψ introduced in (1.1) is linear, the Hilbert space $\mathcal{H}$ inherits a bilinear form from the previous scalar product, that we denote, for all f, g in $\mathcal{H}$,

$$< f, g >_T := < \psi(f), \psi(g) >_{\mathrm{proc}} = \sum_{m=1}^{M} \int_0^T \psi_s^{(m)}(f) \psi_s^{(m)}(g) \mathrm{d}s,$$

and the corresponding quadratic form is denoted $\|f\|_T^2$.

The compensator $\Lambda = (\Lambda^{(m)})_{m=1,...,M}$ of $N = (N^{(m)})_{m=1,...,M}$ is finally defined for all t by

$$\Lambda_t^{(m)} = \int_0^t \lambda_s^{(m)} \mathrm{d}s.$$

Section 2 gives our main oracle inequality and the choice of the ℓ_1-weights in the general framework of counting processes. Section 3 provides the fundamental Bernstein-type inequality. Section 4 details the meaning of the oracle inequality in the Poisson and Aalen setups. The probabilistic results needed for the Hawkes processes as well as the interpretation of the oracle inequality in this framework is done in Section 5. Simulations on multivariate Hawkes processes are performed in Section 6. The last Section is dedicated to the proofs of our results.

2. Lasso estimate and oracle inequality

We wish to estimate the true underlying intensity so our main goal consists in estimating the parameter f_0. For this purpose we assume we are given Φ a dictionary of functions (whose cardinality is denoted $|\Phi|$) and we define f_a as a linear combination of the functions of Φ, that is,

$$f_a := \sum_{\varphi \in \Phi} a_\varphi \varphi,$$

where $a = (a_\varphi)_{\varphi \in \Phi}$ belongs to $\mathbb{R}^\Phi$. Then, since ψ is linear, we get

$$\psi(f_a) = \sum_{\varphi \in \Phi} a_\varphi \psi(\varphi).$$

We use the following least-squares contrast $\mathcal{C}$ defined on $\mathcal{H}$ by

$$\mathcal{C}(f) := -2\ \psi(f) \bullet N_T + \|f\|_T^2, \quad \forall\, f \in \mathcal{H}. \tag{2.1}$$

This contrast, or some variations of $\mathcal{C}$, have already been used in particular setups (see for instance [52] or [29]). The main heuristic justification lies in following arguments. Since $\psi(f)$ is a predictable process, the compensator at time T of $\mathcal{C}(f)$ is given by

$$\tilde{\mathcal{C}}(f) := -2\psi(f) \bullet \Lambda_T + \|f\|_T^2,$$

which can also be written as $\tilde{\mathcal{C}}(f) = -2 < \psi(f), \lambda >_{\text{proc}} + \|\psi(f)\|_{\text{proc}}^2$. Note that $\tilde{\mathcal{C}}$ is minimum when $\|\psi(f) - \lambda\|_{\text{proc}}$ is minimum. If $\lambda = \psi(f_0)$ and if $\|.\|_T$ is a norm on the Hilbert space $\mathcal{H}$ then the unique minimizer of $\tilde{\mathcal{C}}$ is f_0. Therefore, to get the best linear approximation of λ of the form $\psi(f)$, it is natural to look at minimizers of $\mathcal{C}(f)$. Restricting to linear combinations of functions of Φ, since ψ is linear, we obtain

$$\mathcal{C}(f_a) = -2a'b + a'Ga$$

where a' denotes the transpose of the vector a and for $\varphi_1, \varphi_2 \in \Phi$,

$$b_{\varphi_1} = \psi(\varphi_1) \bullet N_T, \quad G_{\varphi_1,\varphi_2} =< \varphi_1, \varphi_2 >_T . \tag{2.2}$$

Note that both the vector b of dimension $|\Phi|$ and the Gram matrix G of dimensions $|\Phi| \times |\Phi|$ are random but nevertheless observable.

To estimate a we minimize the contrast, $\mathcal{C}(f_a)$, subject to an ℓ_1-penalization on the a-vector. That is, we introduce the following ℓ_1-penalized estimator

$$\hat{a} \in \operatorname{argmin}_{a \in \mathbb{R}^\Phi} \{-2a'b + a'Ga + 2d'|a|\} \tag{2.3}$$

where $|a| = (|a_\varphi|)_{\varphi \in \Phi}$ and $d \in \mathbb{R}_+^\Phi$. With a good choice of d the solution of (2.3) will achieve both sparsity and good statistical properties. Finally, we let $\hat{f} = f_{\hat{a}}$ denote the Lasso estimate associated with $\hat{a}$.

Our first result establishes theoretical properties of $\hat{f}$ by using the classical oracle approach. More precisely, we establish a bound on the risk of $\hat{f}$ if some conditions are true. This is a non-probabilistic result that only relies on the definition of $\hat{a}$ by (2.3). In the next section we will deal with the probabilistic aspect, which is to prove that the conditions are fulfilled with large probability.

Theorem 1. *Let $c > 0$. If*

$$G \succeq cI \tag{2.4}$$

and if for all $\varphi \in \Phi$

$$|b_\varphi - \bar{b}_\varphi| \leq d_\varphi, \tag{2.5}$$

where

$$\bar{b}_\varphi = \psi(\varphi) \bullet \Lambda_T,$$

then there exists an absolute constant C, independent of c, such that

$$\|\psi(\hat{f}) - \lambda\|^2_{proc} \leq C \inf_{a\in\mathbb{R}^\Phi} \left\{ \|\lambda - \psi(f_a)\|^2_{proc} + c^{-1} \sum_{\varphi\in S(a)} d^2_\varphi \right\}, \tag{2.6}$$

where $S(a)$ is the support of a. If $\lambda = \psi(f_0)$, the oracle inequality (2.6) can also be rewritten as

$$\|\hat{f} - f_0\|^2_T \leq C \inf_{a\in\mathbb{R}^\Phi} \left\{ \|f_0 - f_a\|^2_T + c^{-1} \sum_{\varphi\in S(a)} d^2_\varphi \right\}, \tag{2.7}$$

The proof of Theorem 1 is given in Section 7.1. Note that Assumption (2.4) ensures that G is invertible and then coordinates of $\hat{a}$ are finite almost surely. Assumption (2.4) also ensures that $\|f\|_T$ is a real norm on f at least when f is a linear combination of the functions of Φ.

Two terms are involved on the right hand sides of (2.6) and (2.7). The first one is an approximation term and the second one can be viewed as a variance term providing a control of the random fluctuations of the b_φ's around the $\bar{b}_\varphi$'s. Note that $b_\varphi - \bar{b}_\varphi = \psi(\varphi) \bullet (N - \Lambda)_T$ is a martingale (see also the comments after Theorem 2 for more details). The approximation term can be small but the price to pay may be a large support of a, leading to large values for the second term. Conversely, a sparse a leads to a small second term. But in this case the approximation term is potentially larger. Note that if the function f_0 can be approximated by a sparse linear combination of the functions of Φ, then we obtain a sharp control of $\|\hat{f} - f_0\|^2_T$. In particular, if f_0 can be decomposed on the dictionary, so we can write $f_0 = f_{a_0}$ for some $a_0 \in \mathbb{R}^\Phi$, then (2.7) gives

$$\|\hat{f} - f_0\|^2_T \leq Cc^{-1} \sum_{\varphi\in S(a_0)} d^2_\varphi.$$

In this case, the right hand side can be viewed as the sum of the estimation errors made by estimating the components of a_0.

Such oracle inequalities are now classical in the huge literature of Lasso procedures. See for instance [5, 6, 15, 16, 17, 18, 37, 60], who established oracle inequalities in the same spirit as in Theorem 1. We bring out the paper [19], which gives technical and heuristic arguments for justifying optimality of such oracle inequalities (see Section 1.3 of [19]). Most of these papers deal with independent data.

In the sequel, we prove that Assumption (2.4) is satisfied with large probability by using the same approach as [55, 56] and to a lesser extent as Section 2.1 of [19] or [54]. Section 5 is in particular mainly devoted to show that (2.4) holds with large probability for the multivariate Hawkes processes.

For Theorem 1 to be of interest, the condition on the martingale, condition (2.5), needs to hold with large probability as well. From this control, we deduce convenient data-driven ℓ_1-weights that are the key parameters of our estimate. Note that our estimation procedure does not depend on the value of c in (2.4). So knowing the latter is not necessary for implementing our procedure. Therefore, one of the main contributions of the paper is to provide new sharp concentration inequalities that are satisfied by multivariate point processes. This is the main goal of Theorem 3 in Section 3 where we establish Bernstein type inequalities for martingales. We apply it to the control of (2.5). This allows us to derive the following result, which specifies the choice of the d_φ's needed to obtain the oracle inequality with large probability.

Theorem 2. *Let $N = (N^{(m)})_{m=1,\dots,M}$ be a multivariate counting process with predictable intensities $\lambda_t^{(m)}$ and almost surely finite corresponding compensator $\Lambda_t^{(m)}$. Define*

$$\Omega_{V,B} = \left\{ \text{for any } \varphi \in \Phi, \sup_{t\in[0,T],m} |\psi_t^{(m)}(\varphi)| \le B_\varphi \text{ and } (\psi(\varphi))^2 \bullet N_T \le V_\varphi \right\},$$

for positive deterministic constants B_φ and V_φ and

$$\Omega_c = \{G \succeq cI\}.$$

Let x and ε be strictly positive constants and define for all $\varphi \in \Phi$,

$$d_\varphi = \sqrt{2(1+\varepsilon)\hat{V}_\varphi^\mu x} + \frac{B_\varphi x}{3}, \tag{2.8}$$

with

$$\hat{V}_\varphi^\mu = \frac{\mu}{\mu - \phi(\mu)} \, (\psi(\varphi))^2 \bullet N_T + \frac{B_\varphi^2 x}{\mu - \phi(\mu)}$$

for a real number μ such that $\mu > \phi(\mu)$, where $\phi(u) = \exp(u) - u - 1$. Let us consider the Lasso estimator $\hat{f}$ of f_0 defined in Section 2. Then, with probability larger than

$$1 - 4\sum_{\varphi\in\Phi} \left(\frac{\log\left(1 + \frac{\mu V_\varphi}{B_\varphi^2 x}\right)}{\log(1+\varepsilon)} + 1 \right) e^{-x} - \mathbb{P}(\Omega_{V,B}^c) - \mathbb{P}(\Omega_c^c),$$

inequality (2.7) is satisfied, i.e.

$$\|\psi(\hat{f}) - \lambda\|_{proc}^2 \leq C \inf_{a \in \mathbb{R}^\Phi} \left\{ \|\lambda - \psi(f_a)\|_{proc}^2 + c^{-1} \sum_{\varphi \in S(a)} d_\varphi^2 \right\}.$$

If moreover $\lambda = \psi(f_0)$, *then*

$$\|\hat{f} - f_0\|_T^2 \leq C \inf_{a \in \mathbb{R}^\Phi} \left\{ \|f_0 - f_a\|_T^2 + c^{-1} \sum_{\varphi \in S(a)} d_\varphi^2 \right\},$$

where C *is a constant independent of* c, Φ, T *and* M.

The first oracle inequality gives a control of the difference between the true intensity and $\psi(\hat{f})$. The equality $\lambda = \psi(f_0)$ is not required and we can apply this result, for instance, with $\lambda = (\psi(f_0))_+$.

Of course, the smaller the d_φ's the better the oracle inequality. Therefore, when x increases, the probability bound and the d_φ's increase and we have to realize a compromise to obtain a meaningful oracle inequality on an event with large probability. The choice of x is deeply discussed below, in Sections 4 and 5 for theoretical purposes and in Section 6 for practical purposes.

Let us first discuss more deeply the definition of d_φ (derived from subsequent Theorem 3) which seems intricate. Up to a constant depending on the choice of μ and ε, d_φ is of same order as $\max\left(\sqrt{x(\psi(\varphi))^2 \bullet N_T}, B_\varphi x\right)$. To give more insight on the values of d_φ, let us consider the very special case where for any $m \in \{1, \dots, M\}$ for any s, $\psi_s^{(m)}(\varphi) = c_m 1_{\{s \in A_m\}}$, where c_m is a positive constant and A_m a compact set included into $[0, T]$. In this case, by naturally choosing $B_\varphi = \max_{1 \leq m \leq M} c_m$, we have:

$$\sqrt{x(\psi(\varphi))^2 \bullet N_T} \geq B_\varphi x \iff \sum_{m=1}^M c_m^2 N_{A_m}^{(m)} \geq x \max_{1 \leq m \leq M} c_m^2,$$

where $N_{A_m}^{(m)}$ represents the number of points of $N^{(m)}$ falling in A_m. For more general vector functions $\psi(\varphi)$, the term $\sqrt{x(\psi(\varphi))^2 \bullet N_T}$ will dominate $B_\varphi x$ if the number of points of the process lying where $\psi(\varphi)$ is large, is significant. In this case, the leading term in d_φ is expected to be the quadratic term $\sqrt{2(1+\varepsilon)\frac{\mu}{\mu - \phi(\mu)} x(\psi(\varphi))^2 \bullet N_T}$ and the linear terms in x can be viewed as residual terms. Furthermore, note that when μ tends to 0,

$$\frac{\mu}{\mu - \phi(\mu)} = 1 + \frac{\mu}{2} + o(\mu), \quad \frac{x}{\mu - \phi(\mu)} \sim \frac{x}{\mu} \to +\infty$$

since $x > 0$. So, if μ and ε tend to 0, the quadratic term tends to $\sqrt{2x(\psi(\varphi))^2 \bullet N_T}$, but the price to pay is the explosion of the linear term in x. In any case, it is possible to make the quadratic term as close to $\sqrt{2x(\psi(\varphi))^2 \bullet N_T}$ as desired. Basically, this term cannot be improved (see comments after Theorem 3 for probabilistic arguments).

Let us now discuss the choice of x. In more classical contexts such as density estimation based on an n-sample, the choice $x = \gamma \log(n)$ plugged in the parameters analog to the d_φ's is convenient, since it both ensures a small probability bound and a meaningful order of magnitude for the oracle bound (see [5] for instance). See also Sections 4 and 5 for similar evaluations in our setup. But it has also been further established that the choice $\gamma = 1$ is the best. Indeed if the components of d are chosen smaller than the analog of $\sqrt{2x(\psi(\varphi))^2 \bullet N_T}$ in the density framework, then the resulting estimator is definitely bad from the theoretical point of view, but simulations also show that, to some extent, if the components of d are larger than the analog of $\sqrt{2x(\psi(\varphi))^2 \bullet N_T}$, then the estimator deteriorates too. A similar result is out of reach in our setting, but similar conclusions may remain valid here since density estimation often provides some clues about what happens for more intricate heteroscedastic models. In particular, the main heuristic justifying the optimality of this tuning result in the density setting is that the quadratic term (and in particular the constant $\sqrt{2}$) corresponds to the rate of the central limit theorem and in this sense, it provides the "best approximation" for the fluctuations. For further discussion, see the simulation study in Section 6.

Finally, it remains to control $\mathbb{P}(\Omega_{V,B})$ and $\mathbb{P}(\Omega_c)$. These are the goals of Section 4 for Poisson and Aalen models and Section 5 for multivariate Hawkes processes.

3. Bernstein type inequalities for multivariate point processes

We establish a Bernstein type concentration inequality based on boundedness assumptions. This result, which has an interest per se from the probabilistic point of view, is the key result to derive the convenient values for the vector d in Theorem 2 and so is capital from the statistical perspective.

Theorem 3. *Let $N = (N^{(m)})_{m=1,\dots,M}$ be a multivariate counting process with predictable intensities $\lambda_t^{(m)}$ and corresponding compensator $\Lambda_t^{(m)}$ with respect to some given filtration. Let $B > 0$. Let $H = (H^{(m)})_{m=1,\dots,M}$ be a multivariate predictable process such that for all $\xi \in (0, 3)$, for all t,*

$$\exp(\xi H/B) \bullet \Lambda_t < \infty \ a.s. \text{ and } \exp(\xi H^2/B^2) \bullet \Lambda_t < \infty \ a.s. \tag{3.1}$$

Let us consider the martingale defined for all $t \geq 0$ by

$$M_t = H \bullet (N - \Lambda)_t.$$

Let $v > w$ and x be positive constants and let τ be a bounded stopping time. Let us define

$$\hat{V}^\mu = \frac{\mu}{\mu - \phi(\mu)} H^2 \bullet N_\tau + \frac{B^2 x}{\mu - \phi(\mu)}$$

for a real number $\mu \in (0,3)$ such that $\mu > \phi(\mu)$, where $\phi(u) = \exp(u) - u - 1$. Then, for any $\varepsilon > 0$,

$$\mathbb{P}\left(M_\tau \geq \sqrt{2(1+\varepsilon)\hat{V}^\mu x} + \frac{Bx}{3} \text{ and } w \leq \hat{V}^\mu \leq v \text{ and } \sup_{m,t\leq\tau} |H_t^{(m)}| \leq B \right) \leq 2\left(\frac{\log(v/w)}{\log(1+\varepsilon)} + 1 \right) e^{-x}. \quad (3.2)$$

This result is based on the exponential martingale for counting processes, which has been used for a long time in the context of the counting process theory. See for instance [8], [57] or [59]. This basically gives a concentration inequality taking the following form (the result is stated here in its univariate form for comparison purposes): for any $x > 0$,

$$\mathbb{P}\left(M_\tau \geq \sqrt{2\rho x} + \frac{Bx}{3} \text{ and } \int_0^\tau H_s^2 \lambda(s) \mathrm{d}s \leq \rho \text{ and } \sup_{s\in[0,\tau]} |H_s| \leq B \right) \leq e^{-x}. \quad (3.3)$$

In (3.3), ρ is a deterministic upper bound of $v = \int_0^\tau H_s^2 \lambda(s)\mathrm{d}s$, the bracket of the martingale, and consequently the martingale equivalent of the variance term. Moreover B is a deterministic upper bound of $\sup_{s\in[0,\tau]} |H_s|$. The leading term for moderate values of x and τ large enough is consequently $\sqrt{2\rho x}$. The central Limit Theorem for martingales states that, under some assumptions, a sequence of martingales $(M_n)_n$ with respective brackets $(v_n)_n$ tending to a deterministic value $\bar{v}$, once correctly normalized, tends to a Gaussian process with bracket $\bar{v}$. Therefore, a term of the form $\sqrt{2\bar{v}x}$ in the upper bound is not improvable, in particular the constant $\sqrt{2}$. However the replacement of the limit $\bar{v}$ by a deterministic upper bound ρ constitutes a loss. In this sense, Theorem 3 improves the bound and consists of plugging in the unbiased estimate $\hat{v} = \int_0^\tau H_s^2 dN_s$ instead of a non sharp deterministic upper bound of v. Note that we are not able to obtain exactly the term $\sqrt{2}$ but any value strictly larger than $\sqrt{2}$, as close as we want to $\sqrt{2}$ up to some additive terms depending on B that are negligible for moderate values of x.

The proof is based on a peeling argument that was first introduced in [40] for Gaussian processes and is given in Section 7.3.

Note that there exist also inequalities that seem nicer than (3.3) which constitutes the basic brick for our purpose. For instance, in [27], it is established that for any deterministic positive real number θ, for any $x > 0$,

$$\mathbb{P}\left(M_\tau \geq \sqrt{2\theta x} \text{ and } \int_0^\tau H_s^2 \mathrm{d}\Lambda_s + \int_0^\tau H_s^2 \mathrm{d}N_s \leq \theta \right) \leq e^{-x}. \quad (3.4)$$

At first sight, this seems better than Theorem 3 because no linear term depending on B appears, but if we wish to use the estimate $2\int_0^\tau H_s^2 \mathrm{d}N_s$ instead of θ in the inequality, we have to bound $|H_s|$ by some B in any case. Moreover, by doing so, the quadratic term will be of order $\sqrt{4\hat{v}x}$ which is worse than the term $\sqrt{2\hat{v}x}$ derived in Theorem 3, even if this constant $\sqrt{2}$ can only be reached asymptotically in our case.

There exists a better result if the martingale M_t is for instance conditionally symmetric (see [25, 4, 27]): for any $x > 0$,

$$\mathbb{P}\left(M_\tau \geq \sqrt{2\kappa x} \text{ and } \int_0^\tau H_s^2 \mathrm{d}N_s \leq \kappa\right) \leq e^{-x}, \tag{3.5}$$

which seems close to the ideal inequality. But there are actually two major problems with this inequality. First, one needs to assume that the martingale is conditionally symmetric, which cannot be the case in our situation for general counting processes and general dictionaries. Second, it depends on the deterministic upper bound κ instead of $\hat{v}$. To replace κ by $\hat{v}$ and then to apply peeling arguments as in the proof of Theorem 3, we need to assume the existence of a positive constant w such that $\hat{v} \geq w$. But if the process is empty, then $\hat{v} = 0$, so we cannot generally find such a positive lower bound, whereas in our theorem, we can always take $w = \frac{B^2 x}{\mu - \phi(\mu)}$ as a lower bound for $\hat{V}^\mu$.

Finally, note that in Proposition 6 (see Section 7.3), we also derive a similar bound where $\hat{V}^\mu$ is replaced by $\int_0^\tau H_s^2 \mathrm{d}\Lambda_s$. Basically, it means that the same type of results hold for the quadratic characteristic instead of the quadratic variation. Though this result is of little use here, since the quadratic characteristic is not observable, we think that it may be of interest for readers investigating self-normalized results as in [26].

4. Applications to the Poisson and Aalen models

We now apply Theorem 2 to the Poisson and Aalen models. The case of the multivariate Hawkes process, which is much more intricate, will be the subject of the next section.

4.1. The Poisson model

Let us recall that in this case, we observe M i.i.d. Poisson processes with intensity f_0 supported by $[0, 1]$ (with $M \geq 2$) and that the norm is given by $\|f\|^2 = \int_0^1 f^2(x)dx$. We assume that Φ is an orthonormal system for $\|.\|$. In this case,

$$\|.\|_1^2 = M\|.\|^2 \quad \text{and} \quad G = MI,$$

where I is the identity matrix. One applies Theorem 2 with $c = M$ (so $\mathbb{P}(\Omega_c^c) = 0$) and

$$B_\varphi = \|\varphi\|_\infty, \quad V_\varphi = \|\varphi\|_\infty^2 (1 + \delta) M m_1,$$

for $\delta > 0$ and $m_1 = \int_0^1 f_0(t)\mathrm{d}t$. Note that here $T = 1$ and therefore $N_T^{(m)} = N_1^{(m)}$ is the total number of observed points for the mth process. Using

$$\psi(\varphi)^2 \bullet N_1 \leq \|\varphi\|_\infty^2 \sum_{m=1}^M N_1^{(m)}$$

and since the distribution of $\sum_{m=1}^M N_1^{(m)}$ is the Poisson distribution with parameter Mm_1, Cramer-Chernov arguments give:

$$\mathbb{P}(\Omega_{V,B}^c) \leq \mathbb{P}\left(\sum_{m=1}^M N_1^{(m)} > (1+\delta)Mm_1\right) \leq \exp\left(-\{(1+\delta)\ln(1+\delta)-\delta\}Mm_1\right).$$

For $\alpha > 0$, by choosing $x = \alpha\log(M)$, we finally obtain the following corollary derived from Theorem 2.

Corollary 1. *With probability larger than* $1 - C_1\frac{|\Phi|\log(M)}{M^\alpha} - e^{-C_2M}$, *where* C_1 *is a constant depending on* μ, ε, α, δ *and* m_1 *and* C_2 *is a constant depending on* δ *and* m_1, *we have:*

$$\|\hat{f}-f_0\|^2 \leq C\inf_{a\in\mathbb{R}^\Phi}\left\{\|f_0-f_a\|^2 + \frac{1}{M^2}\sum_{\varphi\in S(a)}\left(\log(M)\sum_{m=1}^M\int_0^1\varphi^2(x)\mathrm{d}N_x^{(m)} + \log^2(M)\|\varphi\|_\infty^2\right)\right\},$$

where C *is a constant depending on* μ, ε, α, δ *and* m_1.

To shed some lights on this result, consider an asymptotic perspective by assuming that M is large. Assume also, for sake of simplicity, that f_0 is bounded from below on $[0,1]$. If the dictionary Φ (whose size may depend on M) satisfies

$$\max_{\varphi\in\Phi}\|\varphi\|_\infty = o\left(\sqrt{\frac{M}{\log M}}\right),$$

then, since, almost surely,

$$\frac{1}{M}\sum_{m=1}^M\int_0^1\varphi^2(x)\mathrm{d}N_x^{(m)} \overset{M\to\infty}{\longrightarrow} \int_0^1\varphi^2(x)f_0(x)\mathrm{d}x,$$

almost surely,

$$\begin{aligned}&\frac{1}{M^2}\sum_{\varphi\in S(a)}\left(\log(M)\sum_{m=1}^M\int_0^1\varphi^2(x)\mathrm{d}N_x^{(m)} + \log^2(M)\|\varphi\|_\infty^2\right)\\ &= \frac{\log M}{M}\sum_{\varphi\in S(a)}\int_0^1\varphi^2(x)f_0(x)\mathrm{d}x\times(1+o(1)).\end{aligned}$$

The right hand term corresponds, up to the logarithmic term, to the sum of variance terms when estimating $\int_0^1\varphi(x)f_0(x)\mathrm{d}x$ with $\frac{1}{M}\sum_{m=1}^M\int_0^1\varphi(x)\mathrm{d}N_x^{(m)}$ for $\varphi\in S(a)$. This means that the estimator adaptively achieves the best trade-off between a bias term and a variance term. The logarithmic term is the price to pay for adaptation. Furthermore, when $M\to+\infty$, the inequality of Corollary 1 holds with probability that goes to 1 at a polynomial rate. We refer the reader to [50] for a deep discussion on optimality of such results.

4.2. The Aalen model

Results similar to those presented in this paragraph can be found in [29] under restricted eigenvalues conditions instead of (2.4). Recall that we observe an M-sample $(X^{(m)}, Y^{(m)}, N^{(m)})_{m=1,\dots,M}$, with $Y^{(m)} = (Y_t^{(m)})_{t\in[0,1]}$ and $N^{(m)} = (N_t^{(m)})_{t\in[0,1]}$ (with $M \geq 2$). We assume that $X^{(m)} \in [0,1]$ and that the intensity of $N_t^{(m)}$ is $f_0(t, X^{(m)})Y_t^{(m)}$. We set for any f,

$$\|f\|_e^2 := \mathbb{E}\left(\int_0^1 f^2(t, X^{(1)})\big(Y_t^{(1)}\big)^2 \mathrm{d}t\right).$$

We assume that Φ is an orthonormal system for $\|\cdot\|_2$, the classical norm on $\mathbb{L}_2([0,1]^2)$, and we assume that there exists a positive constant r such that

$$\forall\, f \in \mathbb{L}_2([0,1]^2), \quad \|f\|_e \geq r\|f\|_2, \tag{4.1}$$

so that $\|\cdot\|_e$ is a norm. If we assume, for instance, that the density of the $X^{(m)}$'s is lower bounded by a positive constant c_0 and there exists $c_1 > 0$ such that for any t,

$$\mathbb{E}[(Y_t^{(1)})^2 | X^{(1)}] \geq c_1$$

then (4.1) holds with $r^2 = c_0 c_1$. The empirical version of $\|f\|_e$, denoted $\|f\|_{emp}$, is defined by

$$\|f\|_{emp}^2 := \frac{1}{M}\|f\|_T^2 = \frac{1}{M}\sum_{m=1}^M \int_0^1 f^2(t, X^{(m)})\big(Y_t^{(m)}\big)^2 \mathrm{d}t.$$

Unlike the Poisson model, since the intensity depends on covariates $X^{(m)}$'s and variables $Y^{(m)}$'s, the control of $\mathbb{P}(\Omega_c^c)$ is much more cumbersome for the Aalen case, even if it is less intricate than for Hawkes processes (see Section 5). We have the following result proved in Section 7.5.1.

Proposition 1. *We assume that (4.1) is satisfied, the density of the covariates $X^{(m)}$ is bounded by D and*

$$\sup_{t\in[0,1]} \max_{m\in\{1,\dots,M\}} Y_t^{(m)} \leq 1 \text{ almost surely.} \tag{4.2}$$

We consider an orthonormal dictionary Φ of functions of $\mathbb{L}_2([0,1]^2)$ that may depend on M, and we let r_Φ denote the spectral radius of the matrix $\mathfrak{H}$ whose components are $\mathfrak{H}_{\varphi,\varphi'} = \iint |\varphi(t,x)||\varphi'(t,x)|\mathrm{d}t\mathrm{d}x$. Then, if

$$\max_{\varphi\in\Phi} \|\varphi\|_\infty^2 \times r_\Phi |\Phi| \times \frac{\log M}{M} \to 0, \tag{4.3}$$

when $M \to +\infty$ then, for any $\beta > 0$, there exists $C_1 > 0$ depending on β, D and f_0 such that with $c = C_1 M$, we have

$$\mathbb{P}(\Omega_c^c) = O(|\Phi|^2 M^{-\beta}).$$

Assumption (4.2) is usually satisfied in most of the practical examples where Aalen models are involved. See [2] for explicit examples and see for instance [30, 49] for other articles where this assumption is made. In the sequel, we also assume that there exists a positive constant R such that

$$\max_{m\in\{1,\dots,M\}} N_1^{(m)} \leq R \quad \text{a.s.} \tag{4.4}$$

This assumption, considered by [49], is obviously satisfied in survival analysis where there is at most one death per individual. It could have been relaxed in our setting, by considering exponential moments assumptions, to include Markov cases for instance. However this much simpler assumption allows us to avoid tedious and unnecessary technical aspects since we only wish to illustrate our results in a simple framework. Under (4.2) and (4.4), almost surely,

$$\psi(\varphi)^2 \bullet N_T = \sum_{m=1}^M \int_0^1 [Y_t^{(m)}]^2 \varphi^2(t, X^{(m)}) \mathrm{d}N_t^{(m)} \leq \sum_{m=1}^M \int_0^1 \varphi^2(t, X^{(m)}) \mathrm{d}N_t^{(m)} \leq MR\|\varphi\|_\infty^2.$$

So, we apply Theorem 2 with $B_\varphi = \|\varphi\|_\infty$, $V_\varphi = MR\|\varphi\|_\infty^2$ (so $\mathbb{P}(\Omega_{V,B}) = 1$) and $x = \alpha \log(M)$ for $\alpha > 0$. We finally obtain the following corollary.

Corollary 2. *Assume that (4.2) and (4.4) are satisfied. With probability larger than* $1 - C_1 \frac{|\Phi| \log(M)}{M^\alpha} - \mathbb{P}(\Omega_c^c)$, *where* C_1 *is a constant depending on* μ, ε, α *and* R, *we have:*

$$\|\hat{f} - f_0\|_{emp}^2 \leq C \inf_{a\in\mathbb{R}^\Phi} \Bigg\{ \|f_0 - f_a\|_{emp}^2 + \frac{1}{M^2} \sum_{\varphi\in S(a)} \left(\log(M) \sum_{m=1}^M \int_0^1 \varphi^2(t, X^{(m)}) \mathrm{d}N_t^{(m)} + \log^2(M)\|\varphi\|_\infty^2 \right) \Bigg\}$$

where C *is a constant depending on* μ, ε, α *and* R.

To shed lights on this result, assume that the density of the $X^{(m)}$'s is upper bounded by a constant $\tilde{R}$. In an asymptotic perspective with $M \to \infty$, we have almost surely,

$$\frac{1}{M} \sum_{m=1}^M \int_0^1 \varphi^2(t, X^{(m)}) \mathrm{d}N_t^{(m)} \to \mathbb{E}\left(\int_0^1 \varphi^2(t, X^{(1)}) f_0(t, X^{(1)}) Y^{(1)} \mathrm{d}t \right).$$

But

$$\mathbb{E}\left(\int_0^1 \varphi^2(t, X^{(1)}) f_0(t, X^{(1)}) Y^{(1)} \mathrm{d}t \right) \leq \|f_0\|_\infty \mathbb{E}\left(\int_0^1 \varphi^2(t, X^{(1)}) \mathrm{d}t \right) \leq \tilde{R}\|f_0\|_\infty.$$

So, if the dictionary Φ satisfies

$$\max_{\varphi\in\Phi} \|\varphi\|_\infty = O\left(\sqrt{\frac{M}{\log M}} \right),$$

which is true under (4.3) if $r_\Phi|\Phi| \geq 1$, then, almost surely, the variance term is asymptotically smaller than $\log(M)\frac{|S(a)|\|f_0\|_\infty}{M}$ up to constants. So, we can draw the same conclusions as for the Poisson model. We have not discussed here the choice of Φ and Condition (4.3). This will be extensively done in Section 5.2 where we deal with a similar condition but in a more involved setting.

5. Applications to the case of multivariate Hawkes process

For a multivariate Hawkes model, the parameter $f_0 = (\nu^{(m)}, (h_\ell^{(m)})_{\ell=1,\dots,M})_{m=1,\dots,M}$ belongs to

$$\mathcal{H} = \mathbb{H}^M = \left\{ f = (\mathbf{f}^{(m)})_{m=1,\dots,M} \mid \mathbf{f}^{(m)} \in \mathbb{H} \text{ and } \|f\|^2 = \sum_{m=1}^M \|\mathbf{f}^{(m)}\|^2 \right\}$$

where

$$\mathbb{H} = \Big\{ \mathbf{f} = (\mu, (g_\ell)_{\ell=1,\dots,M}) \mid \mu \in \mathbb{R}\ ,\ g_\ell \text{ with support in } [0,1] \\ \text{and } \|\mathbf{f}\|^2 = \mu^2 + \sum_{\ell=1}^M \int_0^1 g_\ell^2(x)\mathrm{d}x < \infty \Big\}.$$

If one defines the linear predictable transformation κ of $\mathbb{H}$ by

$$\kappa_t(\mathbf{f}) = \mu + \sum_{\ell=1}^M \int_{t-1}^{t-} g_\ell(t-u)\mathrm{d}N_u^{(\ell)}, \tag{5.1}$$

then the transformation ψ on $\mathcal{H}$ is given by

$$\psi_t^{(m)}(f) = \kappa_t(\mathbf{f}^{(m)}).$$

The first oracle inequality of Theorem 2 provides theoretical guaranties of our Lasso methodology in full generality and in particular, even if inhibition takes place (see Section 1.2.3). Since $\Omega_{V,B}$ and Ω_c are observable events, we know whether the oracle inequality holds. However we are not able to determine $\mathbb{P}(\Omega_{V,B})$ and $\mathbb{P}(\Omega_c)$ in the general case. Therefore, in Sections 5.1 and 5.2, we assume that all interaction functions are nonnegative and that there exists f_0 in $\mathcal{H}$ so that for any m and any t,

$$\lambda_t^{(m)} = \psi_t^{(m)}(f_0).$$

We also assume that the process is observed on $[-1, T]$ with $T > 1$.

5.1. Some useful probabilistic results for multivariate Hawkes processes

In this paragraph, we present some particular exponential results and tail controls for Hawkes processes. As far as we know, these results are new: They constitute the generalization of [51] to the multivariate case. In this paper, they are used to control $\mathbb{P}(\Omega_c^c)$ and $\mathbb{P}(\Omega_{V,B}^c)$ but they may be of independent interest.

Since the functions $h_\ell^{(m)}$'s are nonnegative, a cluster representation exists. We can indeed construct the Hawkes process by the Poisson cluster representation (see [24]) as follows:

- Distribute *ancestral points* with marks $\ell = 1, \dots, M$ according to homogeneous Poisson processes with intensities $\nu^{(\ell)}$ on $\mathbb{R}$.
- For each ancestral point, form a cluster of descendant points. More precisely, starting with an ancestral point at time 0 of a certain type, we successively build new generations as Poisson processes with intensity $h_\ell^{(m)}(. - T)$, where T is the parent of type ℓ (the corresponding children being of type m). We will be in the situation where this process becomes extinguished and we denote by H the last children of all generations, which also represents the length of the cluster. Note that the number of descendants is a multitype branching process (and there exists a branching cluster representation (see [9, 24, 34])) with offspring distributions being Poisson variables with means
$$\gamma_{\ell,m} = \int_0^1 h_\ell^{(m)}(t)\mathrm{d}t.$$

The essential part we need is that the expected number of offsprings of type m from a point of type ℓ is $\gamma_{\ell,m}$. With $\Gamma = (\gamma_{\ell,m})_{\ell,m=1,\dots,M}$, the theory of multitype branching processes gives that the clusters are finite almost surely if the spectral radius of Γ is strictly smaller than 1. In this case, there is a stationary version of the Hawkes process by the Poisson cluster representation.

Moreover, if Γ has spectral radius strictly smaller than 1, one can provide a bound on the number of points in a cluster. We denote by $\mathbb{P}_\ell$ the law of the cluster whose ancestral point is of type ℓ, $\mathbb{E}_\ell$ is the corresponding expectation.

The following lemma is very general and holds even if the function $h_\ell^{(m)}$ have infinite support as long as the spectral radius Γ is strictly less than 1.

Lemma 1. *If W denotes the total number of points of any type in the cluster whose ancestral point is of type ℓ then if the spectral radius of Γ is strictly smaller than 1 there exists $\vartheta_\ell > 0$, only depending on ℓ and on Γ, such that*
$$\mathbb{E}_\ell(e^{\vartheta_\ell W}) < \infty.$$

This easily leads to the following result, which provides the existence of the Laplace transform of the total number of points in an arbitrary bounded interval, when the functions $h_\ell^{(m)}$ have bounded support.

Proposition 2. *Let N be a stationary multivariate Hawkes process, with compactly supported nonnegative interactions functions and such that the spectral radius of Γ is strictly smaller than 1. For any $A > 0$, let $N_{[-A,0)}$ be the total number of points of N in $[-A, 0)$, all marks included. Then there exists a constant $\theta > 0$, depending on the distribution of the process and on A, such that*

$$\mathcal{E} := \mathbb{E}(e^{\theta N_{[-A,0)}}) < \infty,$$

which implies that for all positive u

$$\mathbb{P}(N_{[-A,0)} \geq u) \leq \mathcal{E} e^{-\theta u}.$$

Moreover one can strengthen the ergodic theorem in a non-asymptotic way, under the same assumptions.

Proposition 3. *Under the assumptions of Proposition 2, let $A > 0$ and let $Z(N)$ be a function depending on the points of N lying in $[-A, 0)$. Assume that there exist b and η nonnegative constants such that*

$$|Z(N)| \leq b(1 + N^{\eta}_{[-A,0)}),$$

where $N_{[-A,0)}$ represents the total number of points of N in $[-A, 0)$, all marks included. We denote $\mathfrak{S}$ the shift operator, meaning that $Z \circ \mathfrak{S}_t(N)$ depends now in the same way as $Z(N)$ on some points that are now the points of N lying in $[t - A, t)$.

We assume $\mathbb{E}[|Z(N)|] < \infty$ and for short, we denote $\mathbb{E}(Z) = \mathbb{E}[Z(N)]$. Then, for any $\alpha > 0$, there exists a constant $\mathcal{T}(\alpha, \eta, f_0, A) > 1$ such that for $T > \mathcal{T}(\alpha, \eta, f_0, A)$, there exist C_1, C_2, C_3 and C_4 positive constants depending on α, η, A and f_0 such that

$$\mathbb{P}\left(\int_0^T [Z \circ \mathfrak{S}_t(N) - \mathbb{E}(Z)]\mathrm{d}t \geq C_1 \sigma \sqrt{T \log^3(T)} + C_2 b (\log(T))^{2+\eta}\right) \leq \frac{C_4}{T^{\alpha}},$$

with $\sigma^2 = \mathbb{E}([Z(N) - \mathbb{E}(Z)]^2 \mathbb{1}_{N_{[-A,0)} \leq \tilde{\mathcal{N}}})$ and $\tilde{\mathcal{N}} = C_3 \log(T)$.

Finally, to deal with the control of $\mathbb{P}(\Omega_c)$, we shall need the next result. First, we define a quadratic form Q on $\mathbb{H}$ by

$$Q(\mathbf{f}, \mathbf{g}) = \mathbb{E}_{\mathbb{P}}\left(\kappa_1(\mathbf{f})\kappa_1(\mathbf{g})\right) = \mathbb{E}_{\mathbb{P}}\left(\frac{1}{T}\int_0^T \kappa_t(\mathbf{f})\kappa_t(\mathbf{g})\mathrm{d}t\right), \quad \mathbf{f}, \mathbf{g} \in \mathbb{H}. \tag{5.2}$$

We have:

Proposition 4. *Under the assumptions of Proposition 2, if the function parameter f_0 satisfies*

$$\min_{m \in \{1,\dots,M\}} \nu^{(m)} > 0 \quad \textit{and} \quad \max_{l,m \in \{1,\dots,M\}} \sup_{t \in [0,1]} h_\ell^{(m)}(t) < \infty \tag{5.3}$$

then there is a constant $\zeta > 0$ such that for any $\mathbf{f} \in \mathbb{H}$,

$$Q(\mathbf{f}, \mathbf{f}) \geq \zeta ||\mathbf{f}||^2.$$

We are now ready to establish oracle inequalities for multivariate Hawkes processes.

5.2. Lasso for Hawkes processes

In the sequel, we still consider the main assumptions of the previous subsection: We deal with a stationary Hawkes process whose intensity is given by (1.3) such that the spectral radius of Γ is strictly smaller than 1 and (5.3) is satisfied. We recall that the components of Γ are the $\gamma_{\ell,m}$'s with

$$\gamma_{\ell,m} = \int_0^1 h_\ell^{(m)}(t)\mathrm{d}t.$$

One of the main results of this section is to link properties of the dictionary (mainly orthonormality but also more involved assumptions) to properties of G (the control of Ω_c). To do so, let us define for all $f \in \mathcal{H}$,

$$\|f\|_\infty = \max\left\{\max_{m=1,\dots,M} |\mu^{(m)}|,\ \max_{m,\ell=1,\dots,M} \|g_\ell^{(m)}\|_\infty\right\}.$$

Then, let us set $\|\Phi\|_\infty := \max\{\|\varphi\|_\infty, \varphi \in \Phi\}$. The next result is based on the probabilistic results of Section 5.1.

Proposition 5. *Assume that the Hawkes process is stationary, that (5.3) is satisfied and that the spectral radius of Γ is strictly smaller than 1. Let r_Φ be the spectral radius of the matrix $\mathfrak{H}$ defined by*

$$\mathfrak{H} = \left(\sum_m \left[|\mu_\varphi^{(m)}||\mu_\rho^{(m)}| + \sum_{\ell=1}^M \int_0^1 |(g_\varphi)_\ell^{(m)}||(g_\rho)_\ell^{(m)}|(u)\mathrm{d}u\right]\right)_{\varphi,\rho\in\Phi}.$$

Assume that Φ is orthonormal and that

$$A_\Phi(T) := r_\Phi \|\Phi\|_\infty^2 |\Phi| [\log(\|\Phi\|_\infty) + \log(|\Phi|)] \frac{\log^5(T)}{T} \to 0 \tag{5.4}$$

when $T \to \infty$. Then, for any $\beta > 0$, there exists $C_1 > 0$ depending on β and f_0 such that with $c = C_1 T$, we have

$$\mathbb{P}(\Omega_c^c) = O(|\Phi|^2 T^{-\beta}).$$

Up to logarithmic terms, (5.4) is similar to (4.3) with M replaced with T. The dictionary Φ can be built via a dictionary $(\Upsilon_k)_{k=1,\dots,K}$ of functions of $\mathbb{L}_2([0,1])$ (that may depend on T) in the following way. A function $\varphi = (\mu_\varphi^{(m)}, ((g_\varphi)_\ell^{(m)})_\ell)_m$ belongs to Φ if and only if only one of its $M + M^2$ components is non zero and in this case,

- if $\mu_\varphi^{(m)} \neq 0$, then $\mu_\varphi^{(m)} = 1$,
- if $(g_\varphi)_\ell^{(m)} \neq 0$, then there exists $k \in \{1,\dots,K\}$ such that $(g_\varphi)_\ell^{(m)} = \Upsilon_k$.

Note that $|\Phi| = M + KM^2$. Furthermore, assume from now on that $(\Upsilon_k)_{k=1,\dots,K}$ is orthonormal in $\mathbb{L}_2([0,1])$. Then Φ is also orthonormal in $\mathcal{H}$ endowed with $\|.\|$.

Before going further, let us discuss Assumption (5.4). First note that the matrix $\mathfrak{H}$ is block diagonal. The first block is the identity matrix of size M. The other M^2 blocks are identical to the matrix:

$$\mathfrak{H}_K = \left(\int |\Upsilon_{k_1}(u)||\Upsilon_{k_2}(u)| \mathrm{d}u \right)_{1 \le k_1, k_2 \le K}.$$

So, if we denote $\tilde{r}_K$ the spectral radius of $\mathfrak{H}_K$, we have:

$$r_\Phi = \max(1, \tilde{r}_K).$$

We analyze the behavior of $\tilde{r}_K$ with respect to K. Note that for any k_1 and any k_2,

$$(\mathfrak{H}_K)_{k_1,k_2} \ge 0.$$

Therefore,

$$\tilde{r}_K \le \sup_{\|x\|_{\ell_1}=1} \|\mathfrak{H}_K x\|_{\ell_1} \le \max_{k_1} \sum_{k_2} (\mathfrak{H}_K)_{k_1,k_2}.$$

We now distinguish three types of orthonormal dictionaries (remember that M is viewed as a constant here):

- Let us consider regular histograms. The basis is composed of the functions $\Upsilon_k = \delta^{-1/2} \mathbb{1}_{((k-1)\delta, k\delta]}$ with $K\delta = 1$. Therefore $\|\Phi\|_\infty = \delta^{-1/2} = \sqrt{K}$. But $\mathfrak{H}_K$ is the identity matrix and $\tilde{r}_K = 1$. Hence (5.4) is satisfied as soon as

$$\frac{K^2 \log(K) \log^5(T)}{T} \to 0$$

  when $T \to \infty$, which is satisfied if $K = o\left(\frac{\sqrt{T}}{\log^3(T)}\right)$.
- Assume that $\|\Phi\|_\infty$ is bounded by an absolute constant (Fourier dictionaries satisfy this assumption). Since $\tilde{r}_K \le K$, (5.4) is satisfied as soon as

$$\frac{K^2 \log(K) \log^5(T)}{T} \to 0$$

  when $T \to \infty$, which is satisfied if $K = o\left(\frac{\sqrt{T}}{\log^3(T)}\right)$.
- Assume that $(\Upsilon_k)_{k=1,\dots,K}$ is a compactly supported wavelet dictionary where resolution levels belong to the set $\{0, 1, \dots, J\}$. In this case, K is of the same order as 2^J, $\|\Phi\|_\infty$ is of the same order as $2^{J/2}$ and it can be established that $\tilde{r}_K \le C 2^{J/2}$ where C is a constant only depending on the choice of the wavelet system (see [33] for further details). Then, (5.4) is satisfied as soon as

$$\frac{K^{5/2} \log(K) \log^5(T)}{T} \to 0$$

  when $T \to \infty$, which is satisfied if $K = o\left(\frac{T^{2/5}}{\log^{12/5}(T)}\right)$.

To apply Theorem 2, it remains to control $\Omega_{V,B}$. Note that

$$\psi_t^{(m)}(\varphi) = \begin{cases} 1 & \text{if } \mu_\varphi^{(m)} = 1 \\ \int_{t-1}^{t-} \Upsilon_k(t-u)\mathrm{d}N_u^{(\ell)} & \text{if } (g_\varphi)_\ell^{(m)} = \Upsilon_k. \end{cases}$$

Let us define

$$\Omega_{\mathcal{N}} = \left\{ \text{for all } t \in [0,T], \text{for all } m \in \{1,\dots,M\} \text{ we have } N_{[t-1,t]}^{(m)} \leq \mathcal{N} \right\}.$$

We therefore set

$$B_\varphi = 1 \text{ if } \mu_\varphi^{(m)} = 1 \text{ and } B_\varphi = \|\Upsilon_k\|_\infty \mathcal{N} \text{ if } (g_\varphi)_\ell^{(m)} = \Upsilon_k. \tag{5.5}$$

Note that on $\Omega_{\mathcal{N}}$, for any $\varphi \in \Phi$,

$$\sup_{t\in[0,T],m} |\psi_t^{(m)}(\varphi)| \leq B_\varphi.$$

Now, for each $\varphi \in \Phi$, let us determine V_φ that constitutes an upper bound of

$$M_\varphi = \sum_{m=1}^{M} \int_0^T [\psi_t^{(m)}(\varphi)]^2 \mathrm{d}N_t^{(m)}.$$

Note that only one term in this sum is non-zero. We set

$$V_\varphi = \lceil T \rceil \mathcal{N} \text{ if } \mu_\varphi^{(m)} = 1 \text{ and } V_\varphi = \|\Upsilon_k\|_\infty^2 \lceil T \rceil \mathcal{N}^3 \text{ if } (g_\varphi)_\ell^{(m)} = \Upsilon_k, \tag{5.6}$$

where $\lceil T \rceil$ denotes the smallest integer larger than T. With this choice, one has that $\Omega_{\mathcal{N}} \subset \Omega_{V,B}$, which leads to the following result.

Corollary 3. *Assume that the Hawkes process is stationary, that (5.3) is satisfied and that the spectral radius of Γ is strictly smaller than 1. With the choices (5.5) and (5.6),*

$$\mathbb{P}(\Omega_{V,B}) \geq \mathbb{P}(\Omega_{\mathcal{N}}) \geq 1 - C_1 T \exp(-C_2 \mathcal{N}),$$

where C_1 and C_2 are positive constants depending on f_0.

If $\mathcal{N} \gg \log(T)$, then for all $\beta > 0$,

$$\mathbb{P}(\Omega_{V,B}^c) \leq \mathbb{P}(\Omega_{\mathcal{N}}^c) = o(T^{-\beta}).$$

We are now ready to apply Theorem 2.

Corollary 4. *Assume that the Hawkes process is stationary, that (5.3) is satisfied and that the spectral radius of Γ is strictly smaller than 1. Assume that the dictionary Φ is built as previously from an orthonormal family $(\Upsilon_k)_{k=1,\dots,K}$. With the notations of*

Theorem 2, let B_φ be defined by (5.5) and d_φ be defined accordingly with $x = \alpha \log(T)$. Then, with probability larger than

$$1 - 4(M + M^2K)\left(\frac{\log\left(1 + \frac{\mu\lceil T\rceil\mathcal{N}}{\alpha\log(T)}\right)}{\log(1+\varepsilon)} + 1\right)T^{-\alpha} - \mathbb{P}(\Omega_{\mathcal{N}}^c) - \mathbb{P}(\Omega_c^c),$$

$$\frac{1}{T}\|\hat{f} - f_0\|_T^2 \leq C \inf_{a\in\mathbb{R}^\Phi}\left\{\frac{1}{T}\|f_0 - f_a\|_T^2 + \sum_{\varphi\in S(a)}\left(\frac{\log(T)(\psi(\varphi))^2 \bullet N_T}{T^2} + \frac{B_\varphi^2\log^2(T)}{T^2}\right)\right\},$$

where C is a constant depending on f_0, μ, ε, and α.

From an asymptotic point of view, if the dictionary also satisfies (5.4), and if $\mathcal{N} = \log^2(T)$ in (5.5), then for T large enough with probability larger than $1 - C_1 K\log(T)T^{-\alpha}$

$$\frac{1}{T}\|\hat{f} - f_0\|_T^2 \leq C_2 \inf_{a\in\mathbb{R}^\Phi}\left\{\frac{1}{T}\|f_0 - f_a\|_T^2 + \frac{\log^3(T)}{T}\sum_{\varphi\in S(a)}\left[\frac{1}{T}\|\varphi\|_T^2 + \frac{\log^{7/2}(T)}{\sqrt{T}}\|\Phi\|_\infty^2\right]\right\},$$

where C_1 and C_2 are constants depending on M, f_0, μ, ε, and α.

We express the oracle inequality by using $\frac{1}{T}\|.\|_T$ simply because, when T goes to $+\infty$, by ergodicity of the process (see for instance [24], and Proposition 3 for a non asymptotic statement),

$$\frac{1}{T}\|f\|_T^2 = \sum_{m=1}^M \frac{1}{T}\int_0^T (\kappa_t(\mathbf{f}^{(m)}))^2 \mathrm{d}t \longrightarrow \sum_{m=1}^M Q(\mathbf{f}^{(m)}, \mathbf{f}^{(m)})$$

under assumptions of Proposition 5. Note that the right hand side is a true norm on $\mathcal{H}$ by Proposition 4. Note also that

$$\frac{\log^{7/2}(T)}{\sqrt{T}}\|\Phi\|_\infty^2 \overset{T\to\infty}{\to} 0,$$

as soon as (5.4) is satisfied for the Fourier basis and compactly supported wavelets. It is also the case for histograms as soon as $K = o\left(\frac{\sqrt{T}}{\log^{7/2}(T)}\right)$. Therefore, this term can be viewed as a residual one. In those cases, the last inequality can be rewritten as

$$\frac{1}{T}\|\hat{f} - f_0\|_T^2 \leq C \inf_{a\in\mathbb{R}^\Phi}\left\{\frac{1}{T}\|f_0 - f_a\|_T^2 + \frac{\log^3(T)}{T}\sum_{\varphi\in S(a)}\frac{1}{T}\|\varphi\|_T^2\right\},$$

for a different constant C, the probability of this event tending to 1 as soon as $\alpha \geq 1/2$ in the Fourier and histogram cases and $\alpha \geq 2/5$ for the compactly supported wavelet basis. Once again, as mentioned for the Poisson or Aalen models, the right hand side corresponds to a classical "bias-variance" trade off and we obtain a classical oracle inequality up to the logarithmic terms. Note that asymptotics is now with respect to T and not with respect to M as for Poisson or Aalen models. So, the same result, namely Theorem 2, allows to consider both asymptotics.

6. Simulations for the multivariate Hawkes process

This section is devoted to illustrations of our procedure on simulated data of multivariate Hawkes processes and comparisons with the well-known adaptive Lasso procedure proposed by [64]. We consider the general case and we do no longer assume that the functions $h_\ell^{(m)}$ are nonnegative as in Section 5. However, if the parameter $\nu^{(m)}$ is large with respect to the $h_\ell^{(m)}$'s, then $\psi^{(m)}(f_0)$ is nonnegative with large probability and therefore $\lambda^{(m)} = \psi^{(m)}(f_0)$ with large probability. Hence, Theorem 2 implies that $\hat{f}$ is close to f_0.

6.1. Description of the Data

As mentioned in the introduction, Hawkes processes can be used in neuroscience to model the action potentials of individual neurons. So, we perform simulations whose parameters are close, to some extent, to real neuronal data. For a given neuron $m \in \{1, \dots, M\}$, we recall that its activity is modeled by a point process $N^{(m)}$ whose intensity is

$$\lambda_t^{(m)} = \left(\nu^{(m)} + \sum_{\ell=1}^{M} \int_{-\infty}^{t-} h_\ell^{(m)}(t-u)\mathrm{d}N^{(\ell)}(u) \right)_+ .$$

The *interaction function* $h_\ell^{(m)}$ represents the influence of the past activity of the neuron ℓ on the neuron m. The *spontaneous rate* $\nu^{(m)}$ may somehow represent the external excitation linked to all the other neurons that are not recorded. It is consequently of crucial importance not only to correctly infer the interaction functions, but also to reconstruct the spontaneous rates accurately. Usually, activity up to 10 neurons can be recorded in a "stationary" phase during a few seconds (sometimes up to one minute). Typically, the points frequency is of the order of 10-80 Hz and the interaction range between points is of the order of a few milliseconds (up to 20 or 40 ms). We first lead three experiments in the *pure excitation case* where all the interaction functions are nonnegative by simulating multivariate Hawkes processes (two with $M = 2$, one with $M = 8$) based on these typical values. More precisely, we take for any $m \in \{1, \dots, M\}$, $\nu^{(m)} = 20$ and the interaction functions $h_\ell^{(m)}$ are defined as follows (supports of all the functions are assumed to lie in the interval $[0, 0.04]$):

- **Experiment 1: $M = 2$ and piecewise constant functions.**

  $$h_1^{(1)} = 30 \times \mathbb{1}_{(0,0.02]}, \quad h_2^{(1)} = 30 \times \mathbb{1}_{(0,0.01]}, \quad h_1^{(2)} = 30 \times \mathbb{1}_{(0.01,0.02]}, \quad h_2^{(2)} = 0.$$

  In this case, each neuron depends on the other one. The spectral radius of the matrix Γ is 0.725.
- **Experiment 2: $M = 2$ and "smooth" functions.** In this experiment, $h_1^{(1)}$ and $h_1^{(2)}$ are not piecewise constant.

  $$h_1^{(1)}(x) = 100\, e^{-200x} \times \mathbb{1}_{(0,0.04]}(x), \quad h_2^{(1)}(x) = 30 \times \mathbb{1}_{(0,0.02]}(x),$$

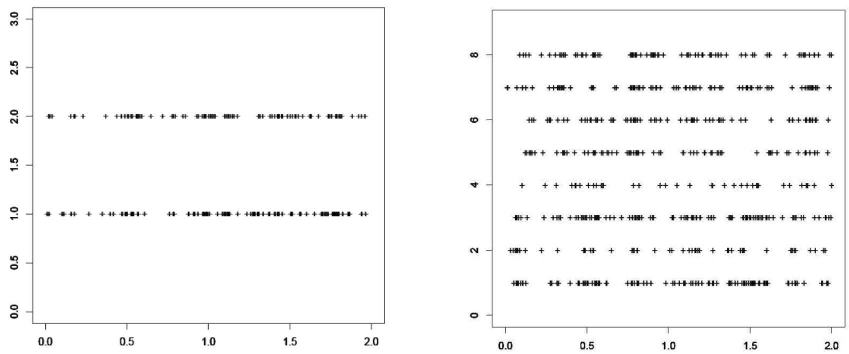

Figure 1. Raster plots of two data sets with $T = 2$ corresponding to Experiment 2 on the left and Experiment 3 on the right. The x-axis correspond to the time of the experiment. Each line with ordinate m corresponds to the points of the process $N^{(m)}$. From bottom to top, we observe 124 and 103 points for Experiment 2 and 101, 60, 117, 38, 73, 75, 86 and 86 points for Experiment 3.

$$h_1^{(2)}(x) = \frac{1}{0.008\sqrt{2\pi}} e^{-\frac{(x-0.02)^2}{2*0.004^2}} \times \mathbb{1}_{(0,0.04]}(x), \quad h_2^{(2)}(x) = 0.$$

In this case, each neuron depends on the other one as well. The spectral radius of the matrix Γ is 0.711.

- **Experiment 3:** $M = 8$ **and piecewise constant functions.**

$$h_2^{(1)} = h_3^{(1)} = h_2^{(2)} = h_1^{(3)} = h_2^{(3)} = h_8^{(5)} = h_5^{(6)} = h_6^{(7)} = h_7^{(8)} = 25 \times \mathbb{1}_{(0,0.02]}$$

and all the other 55 interaction functions are equal to 0. Note in particular that this leads to 3 independent groups of dependent neurons $\{1, 2, 3\}$, $\{4\}$ and $\{5, 6, 7, 8\}$. The spectral radius of the matrix Γ is 0.5.

We also lead one experiment in the *pure inhibition case* where all the interaction functions are nonpositive:

- **Experiment 4:** $M = 2$**.** In this experiment, the interaction functions are the opposite of the functions introduced in Experiment 2. We take for any $m \in \{1, \ldots, M\}$, $\nu^{(m)} = 60$ so that $\psi_t(f_0)$ is positive with high probability.

For each simulation, we let the process "warm up" during 1 second to reach the stationary state[1]. Then the data are collected by taking recordings during the next T seconds. For instance, we record about 100 points per neuron when $T = 2$ and 1000 points when $T = 20$. Figure 1 shows two instances of data sets with $T = 2$.

[1] Note that since the size of the support of the interaction functions is less or equal to 0.04, the "warm up" period is 25 times the interaction range.

6.2. Description of the methods

To avoid approximation errors when computing the matrix G, we focus on a dictionary $(\Upsilon_k)_{k=1,\dots,K}$ whose functions are piecewise constant. More precisely, we take $\Upsilon_k = \delta^{-1/2}\mathbb{1}_{((k-1)\delta,k\delta]}$ with $\delta = 0.04/K$ and K, the size of the dictionary, is chosen later.

Our practical procedure strongly relies on the theoretical one based on the d_φ's defined in (2.8), with x, μ and ε to be specified. First, using Corollary 4, we naturally take $x = \alpha\log(T)$. Then, three hyperparameters would need to be tuned, namely α, μ and ε, if we directly used the Lasso estimate of Theorem 2. So, for simplifications, we implement our procedure by replacing the Lasso parameters $d_\varphi$with

$$\tilde{d}_\varphi(\gamma) = \sqrt{2\gamma\log(T)(\psi(\varphi))^2 \bullet N_T} + \frac{\gamma\log(T)}{3}\sup_{t\in[0,T],m}|\psi_t^{(m)}(\varphi)|,$$

where γ is a constant to be tuned. Besides taking $x = \alpha\log(T)$, our modification consists in neglecting the linear part $\frac{B_\varphi^2 x}{\mu-\phi(\mu)}$ in $\hat{V}^\mu$ and replacing B_φ with $\sup_{t\in[0,T],m}|\psi_t^{(m)}(\varphi)|$. Then, note that, up to these modifications, the choice $\gamma = 1$ corresponds to the limit case where $\alpha \to 1$, $\varepsilon \to 0$ and $\mu \to 0$ in the definition of the d_φ's (see the comments after Theorem 2). Note also that, under the slight abuse consisting in identifying B_φ with $\sup_{t\in[0,T],m}|\psi_t^{(m)}(\varphi)|$, for every parameter μ, ε and α of Theorem 2 with $x = \alpha\ln(T)$, one can find two parameters γ and γ' such that

$$\tilde{d}_\varphi(\gamma) \le d_\varphi \le \tilde{d}_\varphi(\gamma').$$

Therefore, this practical choice is consistent with the theory and tuning hyperparameters reduces to only tuning γ. Our simulation study will provide sound answers to the question of tuning γ.

We compute the Lasso estimate by using the shooting method of [28] and the `R`-package `Lassoshooting`. In particular, we need to invert the matrix G. In all simulations, this matrix was invertible, which is consistent with the fact that Ω_c happens with large probability. Note also that the value of c, namely the smallest eigenvalue of G, can be very small (about 10^{-4}) whereas the largest eigenvalue is potentially as large as 10^5, both values highly depending on the simulation and on T. Fortunately, those values are not needed to compute our Lasso estimate. Since it is based on *Bernstein type inequalities*, our Lasso method is denoted **B** in the sequel.

Due to their soft thresholding nature, Lasso methods are known to underestimate the coefficients [42, 64]. To overcome biases in estimation due to shrinkage, we propose a two steps procedure, as usually suggested in the literature: Once the support of the vector has been estimated by **B**, we compute the ordinary least-square estimator among the vectors a having the same support, which provides the final estimate. This method is denoted **BO** in the sequel.

Another popular method is *adaptive Lasso* proposed by Zou [64]. This method overcomes the flaws of standard Lasso by taking ℓ_1-weights of the form

$$d_\varphi^a(\gamma) = \frac{\gamma}{2|\hat{a}_\varphi^o|^p},$$

where $p > 0$, $\gamma > 0$ and $\hat{a}^o_\varphi$ is a preliminary consistent estimate of the true coefficient. Even if the shapes of the weights are different, the latter are data-driven and this method constitutes a natural competitive method with ours. The most usual choice, which is adopted in the sequel, consists in taking $p = 1$ and the ordinary least squares estimate for the preliminary estimate (see [35, 61, 64]). Then, penalization is stronger for coefficients that are preliminary estimated by small values of the ordinary least square estimate. In the literature, the parameter γ of adaptive Lasso is usually tuned by cross-validation, but this does not make sense for Hawkes data that are fully dependent. Therefore, a preliminary study has been performed to provide meaningful values for γ. Results are given in the next section. This adaptive Lasso method is denoted **A** in the sequel and **AO** when combined with ordinary least squares in the same way as for **BO**.

Simulations are performed in R. The computational time is small (merely a few seconds for one estimate even when $M = 8$, $T = 20$ and $K = 8$ on a classical laptop computer), which constitutes a clear improvement with respect to existing adaptive methods for Hawkes processes. For instance, the "Islands" method[2] of [52] is limited to the estimation of one or two dozens of coefficients at most, because of an extreme computational memory cost whereas here when $M = 8$ and $K = 8$, we can easily deal with $M + KM^2 = 520$ coefficients.

6.3. Results

First, we provide in Figure 2 reconstructions by using the OLS estimate on the whole dictionary, which corresponds to the case where all the weights d_φ are null. As expected, reconstructions are not sparse and also bad due to a small signal to noise ratio (remember that $T = 2$).

Now let us consider methods leading to sparsity. A precise study over 100 simulations has been carried out corresponding to Experiments 1 and 3 for which we can precisely check if the support of the vector $\hat{a}$ is the correct one. For each method, we have selected 3 values for the hyperparameter γ based on results of preliminary simulations. Before studying mean squared errors, we investigate the following problems that are stated in order of importance. We wonder whether our procedure can identify:

- **the dependency groups.** Recall that two neurons belong to the same group if and only if they are connected directly or through the intermediary of one or several neurons. This issue is essential from the neurobiological point of view since knowing interactions between two neurons is of capital importance.
- **the non-zero interaction functions $h^{(m)}_\ell$'s and non-zero spontaneous rates $\nu^{(m)}$'s.** For $\ell, m \in \{1, \dots, M\}$, the neuron ℓ has a significative *direct* interaction on neuron m if and only if $h^{(m)}_\ell \neq 0$;
- **the non-zero coefficients of non-zero interaction functions.** This issue is more mathematical. However, it may provide information about the maximal range

[2] This method developed for $M = 1$ could easily be theoretically adapted for larger values of M, but its extreme computational cost prevents us from using it in practice.

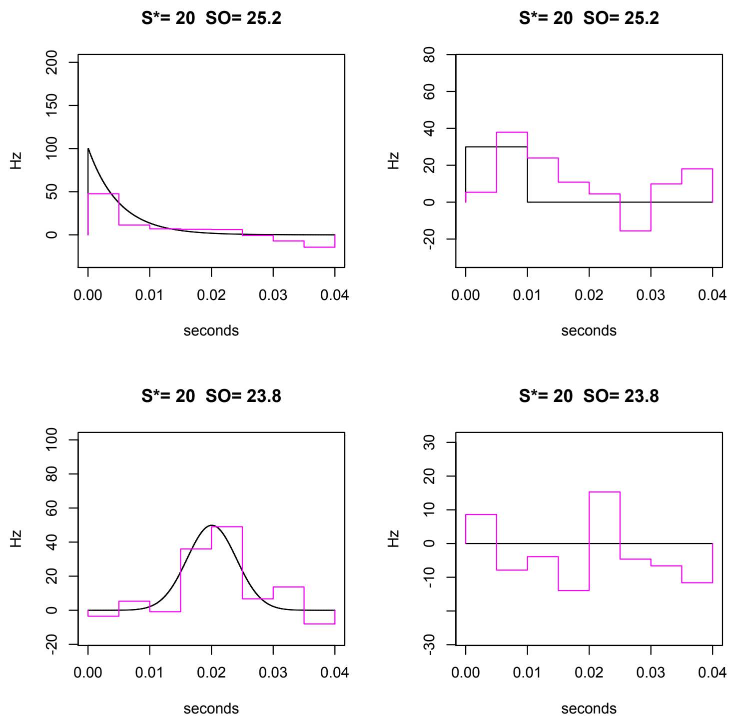

Figure 2. Reconstructions corresponding to Experiment 2 with the OLS estimate with $T = 2$ and $K = 8$. Each line m represents the function $h_\ell^{(m)}$, for $\ell = 1, 2$. The spontaneous rates associated with each line m are given above the graphs where S^* denotes the true spontaneous rate and its estimator is denoted by SO. The true interactions functions are plotted in black whereas the OLS estimates are plotted in magenta.

for direct interactions between two given neurons or about the favored delay of interaction.

Note that the dependency groups are the only features that can be detected by classical analysis tools of neuroscience, such as the Unitary Events method [31]. In particular, to the best of our knowledge, identification of the non-zero interaction functions inside a dependency group is a problem that has not been solved yet as far as we know.

Results for our method and for adaptive Lasso can be found in Table 1. This preliminary study also provides answers for tuning issues. The line "DG" gives the number of correct identifications of dependency groups. For instance, for $M = 8$, "DG" gives the number of simulations for which the 3 dependency groups $\{1, 2, 3\}$, $\{4\}$ and $\{5, 6, 7, 8\}$ are recovered by the methods. When $M = 2$, both methods correctly find that neurons 1 and 2 are dependent, even if $T = 2$. When 8 neurons are considered, the estimates

M=2, T=2	Our Lasso Method			Adaptive Lasso			M=2, T=20	Our Lasso Method			Adaptive Lasso		
γ	0.5	1	2	2	200	1000	γ	0.5	1	2	2	200	1000
DG	**100**	**100**	98	**100**	**100**	98	DG	**100**	**100**	**100**	**100**	**100**	**100**
S	*	*	*	2	2	1	S	*	*	*	*	*	*
F+	**0**	**0**	**0**	1	**0**	**0**	F+	**0**	**0**	**0**	1	**0**	**0**
F-	**0**	**0**	**0**	**0**	**0**	**0**	F-	**0**	**0**	**0**	**0**	**0**	**0**
Coeff+	2	1	**0**	11	2	**0**	Coeff+	1	**0**	**0**	11	2	**0**
Coeff-	**0**	**0**	**0**	**0**	**0**	**0**	Coeff-	**0**	**0**	**0**	**0**	**0**	**0**
SpontMSE	108	140	214	150	193	564	SpontMSE	22	37	69	14	12	27
+ols	104	96	**95**	151	154	516	+ols	11	10	**9**	14	12	10
InterMSE	**7**	9	15	13	8	11	InterMSE	2	3	6	1.4	0.6	0.5
+ols	**7**	**7**	**7**	14	10	10	+ols	0.6	0.5	**0.4**	1.4	0.9	**0.4**

M=8, T=2	Our Lasso Method			Adaptive Lasso			M=8, T=20	Our Lasso Method			Adaptive Lasso		
γ	0.5	1	2	2	200	1000	γ	0.5	1	2	2	200	1000
DG	0	**32**	24	0	0	**32**	DG	63	99	**100**	0	0	90
S	*	*	*	8	7	5	S	*	*	*	*	*	*
F+	17	6	1	55	13	**0.5**	F+	3	1	**0**	55	10	**0**
F-	**0**	**0**	2	**0**	**0**	2	F-	**0**	**0**	**0**	5.5d0	**0**	**0**
Coeff+	22	7	**1**	199.5	17	**1**	Coeff+	4	1	**0**	197	13	**0**
Coeff-	0.5	2	7	**0**	2	7	Coeff-	**0**	**0**	**0**	**0**	**0**	**0**
SpontMSE	**295**	428	768	1445	1026	1835	SpontMSE	82	166	355	104	43	64
+ols	1327	587	859	1512	1058	1935	+ols	41	26	**24**	107	74	26
InterMSE	**38**	51	79	214	49	65	InterMSE	10	19	39	16	2.9	3.17
+ols	63	45	61	228	84	70	+ols	3	2.1	**1.9**	17	6.3	2

Table 1.. Numerical results of both procedures over 100 runs with $K = 4$. Results for Experiment 1 (top) and Experiment 3 (bottom) are given for $T = 2$ (left) and $T = 20$ (right). "DG" gives the number of correct identifications of dependency groups over 100 runs. "S" gives the median number of non-zero spontaneous rate estimates, "*" means that all the spontaneous rate estimates are non-zero over all the simulations. "F+" gives the median number of additional non-zero interaction functions w.r.t. the truth. "F-" gives the median number of missing non-zero interaction functions w.r.t. the truth. "Coeff+" and "Coeff-" are defined in the same way for the coefficients. "SpontMSE" is the Mean Square Error for the spontaneous rates with or without the additional "ordinary least squares step". "InterMSE" is the analog for the interaction functions. In red, we give the optimal values.

should find 3 dependency groups. We see that even with $T = 2$, our method with $\gamma = 1$ correctly guesses the dependency groups for 32% of the simulations. It's close or equal to 100% when $T = 20$ with $\gamma = 1$ or $\gamma = 2$. The adaptive Lasso has to take $\gamma = 1000$ for $T = 2$ and $T = 20$ to obtain as convincing results. Clearly, smaller choices of γ for adaptive Lasso leads to bad estimations of the dependency groups. Next, let us focus on the detection of non-zero spontaneous rates. Whatever the experiment and the parameter γ, our method is optimal whereas adaptive Lasso misses some non-zero spontaneous rates when $T = 2$. Under this criterion, for adaptive Lasso, the choice $\gamma = 1000$ is clearly bad when $T = 2$ (the optimal value of S is $S = 2$ when $M = 2$ and $S = 8$ when $M = 8$) on both experiments, whereas $\gamma = 2$ or $\gamma = 200$ is better. Not surprisingly, the number of additional non-zero functions and additional non-zero coefficients decreases when T grows and when γ grows, whatever the method whereas the number of missing functions or coefficients increases. We can conclude from these facts and from further analysis of Table 1 that the choice $\gamma = 0.5$ for our method and the choice $\gamma = 2$ for the adaptive Lasso are wrong choices of the tuning parameters. In conclusion of this preliminary study, our method with $\gamma = 1$ or $\gamma = 2$ seems a good choice and is robust with respect to T. When $T = 20$, the optimal choice for adaptive Lasso is $\gamma = 1000$. When $T = 2$, the choice is not so clear and depends on the criterion we wish to favor.

Now let us look at mean squared errors (MSE). Since the spontaneous rates do not behave like the other coefficients, we split the MSE in two parts: one for the spontaneous rates:

$$\text{SpontMSE} = \sum_{m=1}^{M} (\hat{\nu}^{(m)} - \nu^{(m)})^2,$$

and one for interactions:

$$\text{InterMSE} = \sum_{m=1}^{M} \sum_{\ell=1}^{M} \int (\hat{h}_\ell^{(m)}(t) - h_\ell^{(m)}(t))^2 \mathrm{d}t.$$

We still report the results for **B**, **BO**, **A** and **AO** in Table 1. Our comments mostly focus on cases where the results for the previous study are good. First, note that results on such cases are better by using the second step (OLS). Furthermore, MSE is increasing with γ for **B** and **A**, since underestimation is stronger when γ increases. This phenomenon does not appear for two step procedures, which leads to a more stable MSE. For adaptive Lasso, when $T = 2$, the choice $\gamma = 200$ leads to good MSE, but the MSE are smaller for **BO** with $\gamma = 1$. When $T = 20$, the choice $\gamma = 1000$ for **AO** leads to results that are of the same magnitude as the ones obtained by **BO** with $\gamma = 1$ or 2. Still for $T = 20$, results for the estimate **B** are worse than results for **A**. It is due to the fact that shrinkage is larger in our method for the coefficients we want to keep than shrinkage of adaptive Lasso that becomes negligible as soon as the true coefficients are large enough. However the second step overcomes this problem.

Note also that a more thorough study of the tuning parameter γ has been performed by [5] where it is mathematically proved that the choice $\gamma < 1$ leads to very degenerate estimates in the density setting. Their method for choosing Lasso parameters being

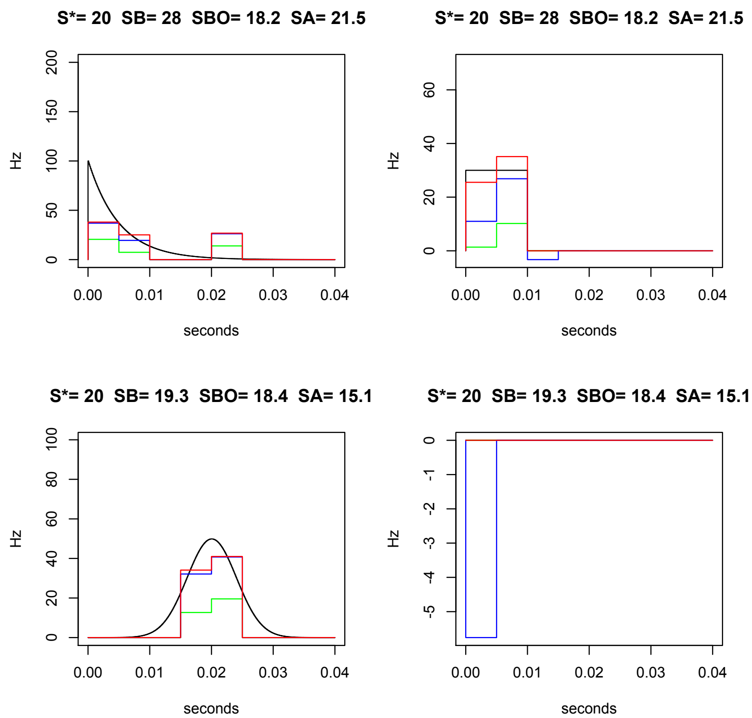

Figure 3. Reconstructions corresponding to Experiment 2 with $T = 2$ and $K = 8$. Each line m represents the function $h_\ell^{(m)}$, for $\ell = 1, 2$. The spontaneous rates estimation associated with each line m is given above the graphs: S* denotes the true spontaneous rate and its estimators computed by using **B**, **BO** and **A** respectively are denoted by SB, SBO and SA. The true interactions functions (in black) are reconstructed by using **B**, **BO** and **A** providing reconstructions in green, red and blue respectively. We use $\gamma = 1$ for **B** and **BO** and $\gamma = 200$ for **A**.

analogous to ours, it seems coherent to obtain worse MSE for $\gamma = 0.5$ than for $\gamma = 1$ or $\gamma = 2$, at least for **BO**. The boundary $\gamma = 1$ in their simulation study seems to be a robust choice there, and it seems to be the case here too.

We now provide some reconstructions by using Lasso methods. Figures 3 and 4 give the reconstructions corresponding to Experiment 2 ($M = 2$) with $K = 8$ for $T = 2$ and $T = 20$ respectively. The reconstructions are quite satisfying. Of course, the quality improves when T grows. We also note improvements by using **BO** instead of **B**. For adaptive Lasso, improvements by using the second step are not significative and this is the reason why we do not represent reconstructions with **AO**. Graphs of the right hand side of Figure 3 illustrate the difficulties of adaptive Lasso to recover the exact support of interactions functions, namely $h_2^{(1)}$ and $h_2^{(2)}$ for $T = 2$. Figure 5 provides another illustration in the case of Experiment 3 ($M = 8$) with $K = 8$ for $T = 20$. For

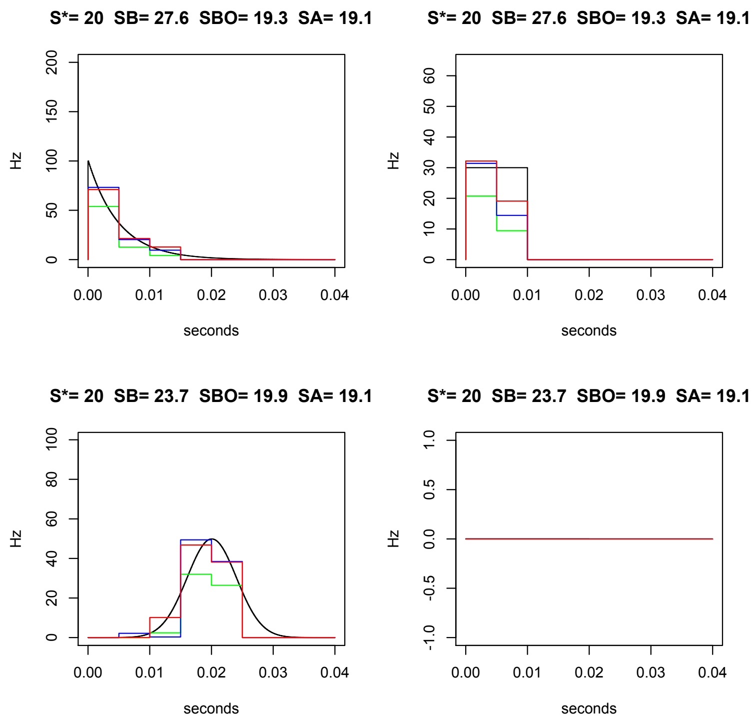

Figure 4. Reconstructions corresponding to Experiment 2 with $T = 20$ and $K = 8$. Same convention as in Figure 3. We use $\gamma = 1$ for **B** and **BO** and $\gamma = 1000$ for **A**.

the sake of clarity, we only represent reconstructions for the first 4 neurons. From the estimation point of view, this illustration provides a clear hierarchy between the methods: **BO** seems to achieve the best results and **B** the worst. Finally, Figure 6 shows that even in the inhibition case, we are able to recover the negative interactions.

6.4. Conclusions

With respect to the problem of tuning our methodology based on Bernstein type inequalities, our simulation study is coherent with theoretical aspects since we achieve our best results by taking $\gamma = 1$, which constitutes the limit case of assumptions of Theorem 2. For practical aspects, we recommend the choice $\gamma = 1$ even if $\gamma = 2$ is acceptable. More importantly, this choice is robust with respect to the duration of recordings, which is not the case for adaptive Lasso. Implemented with $\gamma = 1$, our method outperforms adaptive Lasso and it is able to recover the dependency groups, the non-zero spontaneous rates, the non-zero functions and even the non-zero coefficients as soon as T is large enough. Most

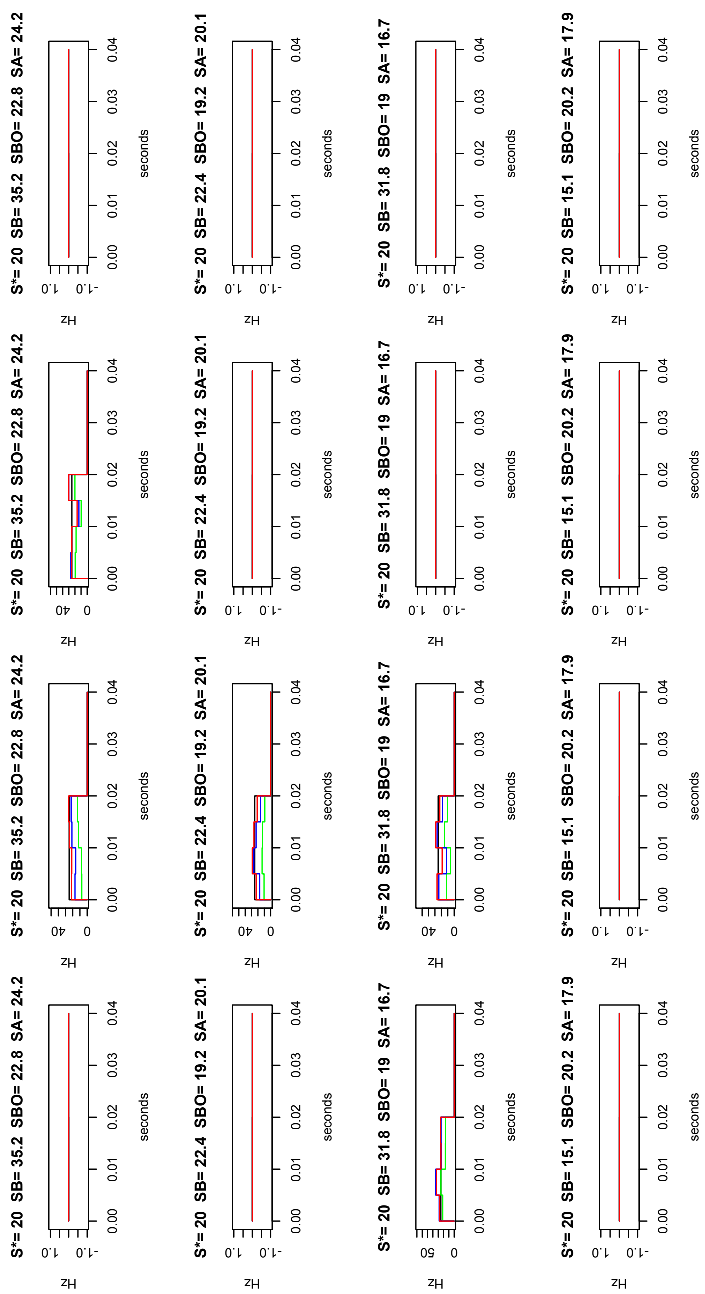

Figure 5. Reconstructions corresponding to Experiment 3 with $T = 20$ and $K = 8$ and for the first 4 neurons. Each line m represents the function $h_\ell^{(m)}$, for $\ell = 1, 2, 3, 4$. Same convention as in Figure 3. We use $\gamma = 1$ for **B** and **BO** and $\gamma = 1000$ for **A**.

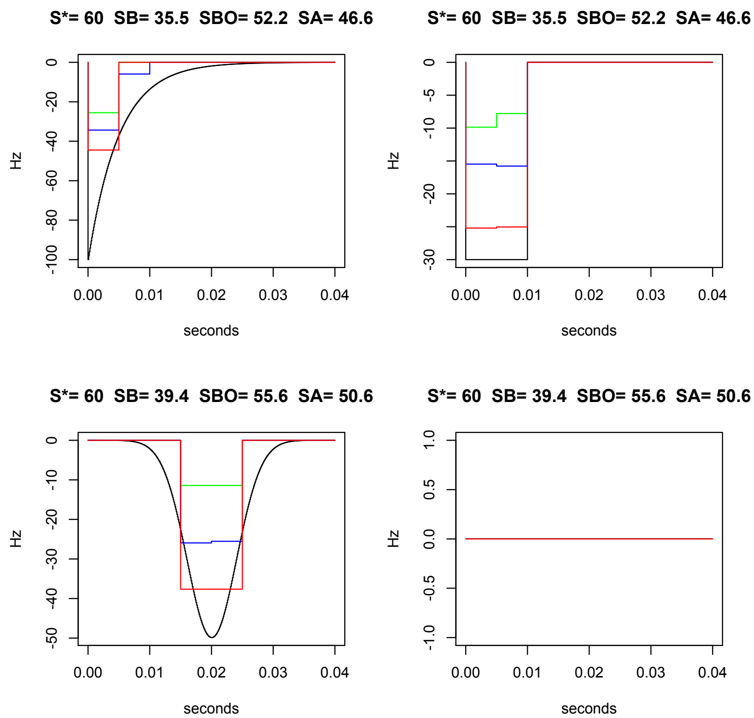

Figure 6. Reconstructions corresponding to Experiment 4 with $T = 20$ and $K = 8$. Same conventions as in Figure 3. We use $\gamma = 1$ for **B** and **BO** and $\gamma = 1000$ for **A**.

of the time, the two step procedure **BO** seems to achieve the best results for parameter estimation.

It is important to note that the question of tuning adaptive Lasso remains open. Some values of γ allow us to obtain very good results but they are not robust with respect to T, which may constitute a serious problem for practitioners. In the standard regression setting, this problem may be overcome by using cross-validation on independent data, which somehow estimates random fluctuations. But in this multivariate Hawkes setup, independence assumptions on data cannot be made and this explains the problems for tuning adaptive Lasso. Our method based on Bernstein type concentration inequalities takes into account those fluctuations. It also takes into account the nature of the coefficients and the variability of their estimates which differ for spontaneous rates on the one hand and coefficients of interaction functions on the other hand. The shape of weights of adaptive Lasso does not incorporate this difference, which explains the contradictions for tuning the method when $T = 2$. For instance, in some cases, adaptive Lasso tends to estimate some spontaneous rates to zero in order to achieve better performance on the interaction functions.

7. Proofs

This section is devoted to the proofs of the results of the paper. Throughout, C is a constant whose value may change from line to line.

7.1. Proof of Theorem 1

The proof of Theorem 1 is standard (see for instance [18]), but for the sake of completeness, we give it. We use $\|.\|_{\ell_2}$ for the Euclidian norm of $\mathbb{R}^\Phi$. Given a recall that

$$f_a = \sum_{\varphi \in \Phi} a_\varphi \varphi.$$

Then, we have $\hat{f} = f_{\hat{a}}$,

$$a'b = \psi(f_a) \bullet N_T$$

and

$$a'Ga = \|f_a\|_T^2 = \|\psi(f_a)\|_{\text{proc}}^2.$$

Then,

$$-2\psi(f_{\hat{a}}) \bullet N_T + \|f_{\hat{a}}\|_T^2 + 2d'|\hat{a}| \leq -2\psi(f_a) \bullet N_T + \|f_a\|_T^2 + 2d'|a|.$$

So,

$$\begin{aligned}
\|\psi(f_{\hat{a}}) - \lambda\|_{\text{proc}}^2 &= \|\psi(f_{\hat{a}})\|_{\text{proc}}^2 + \|\lambda\|_{\text{proc}}^2 - 2 < \psi(f_{\hat{a}}), \lambda >_{\text{proc}} \\
&\leq \|\psi(f_a)\|_{\text{proc}}^2 + \|\lambda\|_{\text{proc}}^2 + 2\psi(f_{\hat{a}} - f_a) \bullet N_T \\
&\qquad\qquad + 2d'\left(|a| - |\hat{a}|\right) - 2 < \psi(f_{\hat{a}}), \lambda >_{\text{proc}} \\
&= \|\psi(f_a) - \lambda\|_{\text{proc}}^2 + 2 < \psi(f_a - f_{\hat{a}}), \lambda >_{\text{proc}} \\
&\qquad\qquad + 2\psi(f_{\hat{a}} - f_a) \bullet N_T + 2d'\left(|a| - |\hat{a}|\right) \\
&= \|\psi(f_a) - \lambda\|_{\text{proc}}^2 + 2\psi(f_a - f_{\hat{a}}) \bullet (\Lambda - N)_T + 2d'\left(|a| - |\hat{a}|\right) \\
&= \|\psi(f_a) - \lambda\|_{\text{proc}}^2 + 2\sum_{\varphi \in \Phi} (a_\varphi - \hat{a}_\varphi)\psi(\varphi) \bullet (\Lambda - N)_T + 2d'\left(|a| - |\hat{a}|\right) \\
&\leq \|\psi(f_a) - \lambda\|_{\text{proc}}^2 + 2\sum_{\varphi \in \Phi} |a_\varphi - \hat{a}_\varphi| \times |\bar{b}_\varphi - b_\varphi| + 2d'\left(|a| - |\hat{a}|\right).
\end{aligned}$$

Using (2.5), we obtain:

$$\begin{aligned}
\|\psi(f_{\hat{a}}) - \lambda\|_{\text{proc}}^2 &\leq \|\psi(f_a) - \lambda\|_{\text{proc}}^2 + 2\sum_{\varphi \in \Phi} d_\varphi |a_\varphi - \hat{a}_\varphi| + 2\sum_{\varphi \in \Phi} d_\varphi \left(|a_\varphi| - |\hat{a}_\varphi|\right) \\
&\leq \|\psi(f_a) - \lambda\|_{\text{proc}}^2 + 2\sum_{\varphi \in \Phi} d_\varphi \left(|a_\varphi - \hat{a}_\varphi| + |a_\varphi| - |\hat{a}_\varphi|\right).
\end{aligned}$$

Now, if $\varphi \notin S(a)$, $|a_\varphi - \hat{a}_\varphi| + |a_\varphi| - |\hat{a}_\varphi| = 0$, and

$$\begin{aligned}
\|\psi(f_{\hat{a}}) - \lambda\|^2_{\mathrm{proc}} &\leq \|\psi(f_a) - \lambda\|^2_{\mathrm{proc}} + 2 \sum_{\varphi \in S(a)} d_\varphi \left(|a_\varphi - \hat{a}_\varphi| + |a_\varphi| - |\hat{a}_\varphi|\right) \\
&\leq \|\psi(f_a) - \lambda\|^2_{\mathrm{proc}} + 4 \sum_{\varphi \in S(a)} d_\varphi \left(|a_\varphi - \hat{a}_\varphi|\right) \\
&\leq \|\psi(f_a) - \lambda\|^2_{\mathrm{proc}} + 4\|\hat{a} - a\|_{\ell_2} \left(\sum_{\varphi \in S(a)} d_\varphi^2\right)^{1/2}.
\end{aligned}$$

We now use the assumption on the Gram matrix given by (2.4) and the triangular inequality for $\|.\|_T$, which yields

$$\begin{aligned}
\|\hat{a} - a\|^2_{\ell_2} &\leq c^{-1} (\hat{a} - a)' G (\hat{a} - a) \\
&= c^{-1} \|f_{\hat{a}} - f_a\|^2_T \\
&\leq 2c^{-1} \left(\|\psi(f_{\hat{a}}) - \lambda\|^2_{\mathrm{proc}} + \|\psi(f_a) - \lambda\|^2_{\mathrm{proc}}\right).
\end{aligned}$$

Let us take $\alpha \in (0,1)$. Since for any $x \in \mathbb{R}$ and any $y \in \mathbb{R}$, $2xy \leq \alpha x^2 + \alpha^{-1} y^2$, we obtain:

$$\begin{aligned}
\|\psi(f_{\hat{a}}) - \lambda\|^2_{\mathrm{proc}} &\leq \|\psi(f_a) - \lambda\|^2_{\mathrm{proc}} \\
&\quad + 4\sqrt{2} c^{-1/2} \sqrt{\|\psi(f_{\hat{a}}) - \lambda\|^2_{\mathrm{proc}} + \|\psi(f_a) - \lambda\|^2_{\mathrm{proc}}} \left(\sum_{\varphi \in S(a)} d_\varphi^2\right)^{1/2} \\
&\leq \|\psi(f_a) - \lambda\|^2_{\mathrm{proc}} \\
&\quad + \alpha \left(\|\psi(f_{\hat{a}}) - \lambda\|^2_{\mathrm{proc}} + \|\psi(f_a) - \lambda\|^2_{\mathrm{proc}}\right) + 8\alpha^{-1} c^{-1} \sum_{\varphi \in S(a)} d_\varphi^2 \\
&\leq (1-\alpha)^{-1} \left((1+\alpha)\|\psi(f_a) - \lambda\|^2_{\mathrm{proc}} + 8\alpha^{-1} c^{-1} \sum_{\varphi \in S(a)} d_\varphi^2\right).
\end{aligned}$$

The theorem is proved just by taking an arbitrary absolute value for $\alpha \in (0,1)$.

7.2. Proof of Theorem 2

Let us first define

$$\mathcal{T} = \{t \geq 0 \ \ / \ \ \sup_m |\psi_t^{(m)}(\varphi)| > B_\varphi\}. \tag{7.1}$$

Let us define the stopping time $\tau' = \inf \mathcal{T}$ and the predictible process H by

$$H_t^{(m)} = \psi_t^{(m)}(\varphi) \mathbb{1}_{t \leq \tau'}.$$

Let us apply Theorem 3 to this choice of H with $\tau = T$ and $B = B_\varphi$. The choice of v and w will be given later on. To apply this result, we need to check that for all t and all

$\xi \in (0,3)$, $\sum_m \int_0^t e^{\xi \frac{H_s^{(m)}}{B_\varphi}} \lambda_s^{(m)} \mathrm{d}s$ is a.s. finite. But if $t > \tau'$, then

$$\int_0^t e^{\xi \frac{H_s^{(m)}}{B_\varphi}} \lambda_s^{(m)} \mathrm{d}s = \int_0^{\tau'} e^{\xi \frac{H_s^{(m)}}{B_\varphi}} \lambda_s^{(m)} \mathrm{d}s + \int_{\tau'}^t \lambda_s^{(m)} \mathrm{d}s,$$

where the second part is obviously finite (it is just $\Lambda_t^{(m)} - \Lambda_{\tau'}^{(m)}$.) Hence it remains to prove that for all $t \leq \tau'$,

$$\int_0^t e^{\xi \frac{H_s^{(m)}}{B_\varphi}} \lambda_s^{(m)} \mathrm{d}s$$

is finite. But for all $s < t$, $s < \tau'$ and consequently $s \notin \mathcal{T}$. Therefore $|H_s^{(m)}| \leq B_\varphi$. Since we are integrating with respect to the Lebesgue measure, the fact that it eventually does not hold in t is not a problem and

$$\int_0^t e^{\xi \frac{H_s^{(m)}}{B_\varphi}} \lambda_s^{(m)} \mathrm{d}s \leq e^{\xi} \Lambda_t^{(m)},$$

which is obviously finite a.s. The same reasoning can be applied to show that a.s. $\exp(\xi H^2/B^2) \bullet \Lambda_t < \infty$. We can also apply Theorem 3 to $-H$ in the same way. We obtain at the end that for all $\varepsilon > 0$

$$\mathbb{P}\left(|H \bullet (N - \Lambda)_T| \geq \sqrt{2(1+\varepsilon)\hat{V}^\mu x} + \frac{B_\varphi x}{3} \text{ and } w \leq \hat{V}^\mu \leq v \text{ and } \sup_{m,t \leq T} |H_t^{(m)}| \leq B_\varphi \right) \leq 4 \left(\frac{\log(v/w)}{\log(1+\varepsilon)} + 1 \right) e^{-x}. \quad (7.2)$$

But on $\Omega_{V,B}$ it is clear that $\forall t \in [0,T], t \notin \mathcal{T}$. Therefore $\tau' \geq T$. Therefore for all $t \leq T$, one also has $t \leq \tau'$ and $H_t^{(m)} = \psi_t^{(m)}(\varphi)$. Consequently, on $\Omega_{V,B}$,

$$H \bullet (N - \Lambda)_T = b_\varphi - \bar{b}_\varphi \text{ and } \hat{V}^\mu = \hat{V}_\varphi^\mu.$$

Moreover, on $\Omega_{V,B}$, one has that

$$\frac{B_\varphi^2 x}{\mu - \phi(\mu)} \leq \hat{V}_\varphi^\mu \leq \frac{\mu}{\mu - \phi(\mu)} V_\varphi + \frac{B_\varphi^2 x}{\mu - \phi(\mu)}.$$

So, we take w and v as respectively the left and right hand side of the previous inequality. Finally note that on $\Omega_{V,B}$,

$$\sup_{m,t \leq T} |H_t^{(m)}| = \sup_{m,t \leq T} |\psi_t^{(m)}(\varphi)| \leq B_\varphi.$$

Hence, we can rewrite (7.2) as follows

$$\mathbb{P}\left(|b_\varphi - \bar{b}_\varphi| \geq \sqrt{2(1+\varepsilon)\hat{V}_\varphi^\mu x} + \frac{B_\varphi x}{3} \text{ and } \Omega_{V,B} \right) \leq 4 \left(\frac{\log\left(1 + \frac{\mu V_\varphi}{B_\varphi^2 x}\right)}{\log(1+\varepsilon)} + 1 \right) e^{-x}. \quad (7.3)$$

Apply this to all $\varphi \in \Phi$, we obtain that

$$\mathbb{P}\left(\exists\, \varphi \in \Phi \text{ s.t. } |b_\varphi - \bar{b}_\varphi| \geq d_\varphi \text{ and } \Omega_{V,B}\right) \leq 4 \sum_{\varphi \in \Phi} \left(\frac{\log\left(1 + \frac{\mu V_\varphi}{B_\varphi^2 x}\right)}{\log(1+\varepsilon)} + 1 \right) e^{-x}.$$

Now on the event $\Omega_c \cap \Omega_{V,B} \cap \{\forall \varphi \in \Phi, |b_\varphi - \bar{b}_\varphi| \leq d_\varphi\}$, one can apply Theorem 1. To obtain Theorem 2, it remains to bound the probability of the complementary event by

$$\mathbb{P}(\Omega_c^c) + \mathbb{P}(\Omega_{V,B}^c) + \mathbb{P}\left(\exists\, \varphi \in \Phi \text{ s.t. } |b_\varphi - \bar{b}_\varphi| \geq d_\varphi \text{ and } \Omega_{V,B}\right).$$

7.3. Proof of Theorem 3

First, replacing H with H/B, we can always assume that $B = 1$. Next, let us fix for the moment $\xi \in (0,3)$. If one assumes that almost surely for all $t > 0$, $\sum_{m=1}^M \int_0^t e^{\xi H_s^{(m)}} \lambda_s^{(m)} \mathrm{d}s < \infty$ (i.e. that the process $e^{\xi H} \bullet \Lambda$ is well defined) then one can apply Theorem 2 of [8, p165], stating that the process $(E_t)_{t \geq 0}$ defined for all t by

$$E_t = \exp(\xi H \bullet (N - \Lambda)_t - \phi(\xi H) \bullet \Lambda_t)$$

is a supermartingale. It is also the case for $E_{t\wedge\tau}$ if τ is a bounded stopping time. Hence for any $\xi \in (0,3)$ and for any $x > 0$, one has that

$$\mathbb{P}(E_{t\wedge\tau} > e^x) \leq e^{-x}\mathbb{E}(E_{t\wedge\tau}) \leq e^{-x},$$

which means that

$$\mathbb{P}(\xi H \bullet (N - \Lambda)_{t\wedge\tau} - \phi(\xi H) \bullet \Lambda_{t\wedge\tau} > x) \leq e^{-x}.$$

Therefore

$$\mathbb{P}(\xi H \bullet (N - \Lambda)_{t\wedge\tau} - \phi(\xi H) \bullet \Lambda_{t\wedge\tau} > x \text{ and } \sup_{s \leq \tau, m} |H_s^{(m)}| \leq 1) \leq e^{-x}.$$

But if $\sup_{s \leq \tau, m} |H_s^{(m)}| \leq 1$, then for any $\xi > 0$ and any s,

$$\phi(\xi H_s^{(m)}) \leq (H_s^{(m)})^2 \phi(\xi).$$

So, for every $\xi \in (0,3)$, we obtain:

$$\mathbb{P}\left(M_\tau \geq \xi^{-1}\phi(\xi) H^2 \bullet \Lambda_\tau + \xi^{-1} x \text{ and } \sup_{s \leq \tau, m} |H_s^{(m)}| \leq 1\right) \leq e^{-x}. \tag{7.4}$$

Now let us focus on the event $H^2 \bullet \Lambda_\tau \leq v$ where v is a deterministic quantity. We have that consequently

$$\mathbb{P}\left(M_\tau \geq \xi^{-1}\phi(\xi) v + \xi^{-1} x \text{ and } H^2 \bullet \Lambda_\tau \leq v \text{ and } \sup_{s \leq \tau, m} |H_s^{(m)}| \leq 1\right) \leq e^{-x}.$$

It remains to choose ξ such that $\xi^{-1}\phi(\xi)v + \xi^{-1}x$ is minimal. But this expression has no simple form. However, since $0 < \xi < 3$, one can bound $\phi(\xi)$ by $\xi^2(1-\xi/3)^{-1}/2$. Hence we can start with

$$\mathbb{P}\left(M_\tau \geq \frac{\xi}{2(1-\xi/3)} H^2 \bullet \Lambda_\tau + \xi^{-1}x \text{ and } \sup_{s\leq\tau,m} |H_s^{(m)}| \leq 1\right) \leq e^{-x} \tag{7.5}$$

and also

$$\mathbb{P}\left(M_\tau \geq \frac{\xi}{2(1-\xi/3)} v + \xi^{-1}x \text{ and } H^2 \bullet \Lambda_\tau \leq v \text{ and } \sup_{s\leq\tau,m} |H_s^{(m)}| \leq 1\right) \leq e^{-x}. \tag{7.6}$$

It remains now to minimize $\xi \longmapsto \frac{\xi}{2(1-\xi/3)} v + \xi^{-1}x$.

Lemma 2. *Let a, b and x be positive constants and let us consider on $(0, 1/b)$,*

$$g(\xi) = \frac{a\xi}{(1-b\xi)} + \frac{x}{\xi}.$$

Then $\min_{\xi\in(0,1/b)} g(\xi) = 2\sqrt{ax} + bx$ and the minimum is achieved in $\xi(a,b,x) = \frac{xb-\sqrt{ax}}{xb^2-a}$.

Proof. The limits of g in 0^+ and $(1/b)^-$ are $+\infty$. The derivative is given by

$$g'(\xi) = \frac{a}{(1-b\xi)^2} - \frac{x}{\xi^2}$$

which is null in $\xi(a,b,x)$ (remark that the other solution of the polynomial does not lie in $(0, 1/b)$). Finally it remains to evaluate the quantity in $\xi(a,b,x)$ to obtain the result. $\square$

Now, we apply (7.6) with $\xi(v/2, 1/3, x)$ and we obtain this well known formula which can be found in [57] for instance:

$$\mathbb{P}\left(M_\tau \geq \sqrt{2vx} + x/3 \text{ and } H^2 \bullet \Lambda_\tau \leq v \text{ and } \sup_{s\leq\tau,m} |H_s^{(m)}| \leq 1\right) \leq e^{-x}. \tag{7.7}$$

Now we would like first to replace v by its random version $H^2 \bullet \Lambda_\tau$. Let w, v be some positive constants and let us concentrate on the event

$$w \leq H^2 \bullet \Lambda_\tau \leq v. \tag{7.8}$$

For all $\varepsilon > 0$ we introduce K a positive integer depending on ε, v and w such that $(1+\varepsilon)^K w \geq v$. Note that $K = \lceil \log(v/w)/\log(1+\varepsilon) \rceil$ is a possible choice. Let us denote $v_0 = w$, $v_1 = (1+\varepsilon)w$, ..., $v_K = (1+\varepsilon)^K w$. For any $0 < \xi < 3$ and any k in $\{0, ..., K-1\}$, one has, by applying (7.5),

$$\mathbb{P}\left(M_\tau \geq \frac{\xi}{2(1-\xi/3)} H^2 \bullet \Lambda_\tau + \xi^{-1}x \text{ and } v_k \leq H^2 \bullet \Lambda_\tau \leq v_{k+1} \text{ and } \sup_{s\leq\tau,m} |H_s^{(m)}| \leq 1\right) \leq e^{-x}.$$

This implies that

$$\mathbb{P}\left(M_\tau \geq \frac{\xi}{2(1-\xi/3)}v_{k+1} + \xi^{-1}x \text{ and } v_k \leq H^2 \bullet \Lambda_\tau \leq v_{k+1} \text{ and } \sup_{s\leq \tau, m} |H_s^{(m)}| \leq 1\right) \leq e^{-x}.$$

Using Lemma 2, with $\xi = \xi(v_{k+1}/2, 1/3, x)$, this gives

$$\mathbb{P}\left(M_\tau \geq \sqrt{2v_{k+1}x} + x/3 \text{ and } v_k \leq H^2 \bullet \Lambda_\tau \leq v_{k+1} \text{ and } \sup_{s\leq \tau, m} |H_s^{(m)}| \leq 1\right) \leq e^{-x}.$$

But if $v_k \leq H^2 \bullet \Lambda_\tau$, $v_{k+1} \leq (1+\varepsilon)v_k \leq (1+\varepsilon)H^2 \bullet \Lambda_\tau$, so

$$\mathbb{P}\left(M_\tau \geq \sqrt{2(1+\varepsilon)\,(H^2 \bullet \Lambda_\tau)\,x} + x/3 \text{ and } v_k \leq H^2 \bullet \Lambda_\tau \leq v_{k+1} \text{ and } \sup_{s\leq \tau, m} |H_s^{(m)}| \leq 1\right) \leq e^{-x}.$$

Finally summing on k, this gives

$$\mathbb{P}\left(M_\tau \geq \sqrt{2(1+\varepsilon)(H^2 \bullet \Lambda_\tau)x} + x/3 \text{ and } w \leq H^2 \bullet \Lambda_\tau \leq v \right.$$
$$\left. \text{and } \sup_{s\leq \tau, m} |H_s^{(m)}| \leq 1\right) \leq Ke^{-x}. \quad (7.9)$$

This leads to the following result that has interest per se.

Proposition 6. *Let $N = (N^{(m)})_{m=1,\dots,M}$ be a multivariate counting process with predictable intensities $\lambda_t^{(m)}$ and corresponding compensator $\Lambda_t^{(m)}$ with respect to some given filtration. Let $B > 0$. Let $H = (H^{(m)})_{m=1,\dots,M}$ be a multivariate predictable process such that for all $\xi \in (0,3)$, $e^{\xi H/B} \bullet \Lambda_t < \infty$ a.s. for all t. Let us consider the martingale defined for all t by*

$$M_t = H \bullet (N - \Lambda)_t.$$

Let $v > w$ be positive constants and let τ be a bounded stopping time. Then for any $\varepsilon, x > 0$

$$\mathbb{P}\left(M_\tau \geq \sqrt{2(1+\varepsilon)(H^2 \bullet \Lambda_\tau)x} + \frac{Bx}{3} \textit{ and } w \leq H^2 \bullet \Lambda_\tau \leq v \right.$$
$$\left. \textit{and } \sup_{m, t\leq \tau} |H_t^{(m)}| \leq B\right) \leq \left(\frac{\log(v/w)}{\log(1+\varepsilon)} + 1\right) e^{-x}. \quad (7.10)$$

Next, we would like to replace $H^2 \bullet \Lambda_\tau$, the quadratic characteristic of M, with its estimator $H^2 \bullet N_\tau$, i.e. the quadratic variation of M. For this purpose, let us consider $W_t = -H^2 \bullet (N - \Lambda)_t$ which is still a martingale since the $-(H_s^{(m)})^2$'s are still predictable processes. We apply (7.4) with μ instead of ξ, noticing that on the event $\{\sup_{s\leq\tau,m} |H_s^{(m)}| \leq 1\}$, one has that $H^4 \bullet \Lambda_\tau \leq H^2 \bullet \Lambda_\tau$. This gives that

$$\mathbb{P}\left(H^2 \bullet \Lambda_\tau \geq H^2 \bullet N_\tau + \{\phi(\mu)/\mu\}H^2 \bullet \Lambda_\tau + x/\mu \text{ and } \sup_{s\leq \tau, m} |H_s^{(m)}| \leq 1\right) \leq e^{-x},$$

which means that

$$\mathbb{P}\left(H^2 \bullet \Lambda_\tau \geq \hat{V}^\mu \text{ and } \sup_{s\leq\tau,m} |H_s^{(m)}| \leq 1\right) \leq e^{-x}. \tag{7.11}$$

So we use again (7.5) combined with (7.11) to obtain that for all $\xi \in (0,3)$

$$\mathbb{P}\left(M_\tau \geq \frac{\xi}{2(1-\xi/3)}\hat{V}^\mu + \xi^{-1}x \text{ and } \sup_{s\leq\tau,m} |H_s^{(m)}| \leq 1\right) \leq$$
$$\mathbb{P}\left(M_\tau \geq \frac{\xi}{2(1-\xi/3)}\hat{V}^\mu + \xi^{-1}x \text{ and } \sup_{s\leq\tau,m} |H_s^{(m)}| \leq 1 \text{ and } H^2 \bullet \Lambda_\tau \leq \hat{V}^\mu\right) +$$
$$+\,\mathbb{P}\left(H^2 \bullet \Lambda_\tau \geq \hat{V}^\mu \text{ and } \sup_{s\leq\tau,m} |H_s^{(m)}| \leq 1\right) \leq 2e^{-x}.$$

This new inequality replaces (7.5) and it remains to replace $H^2 \bullet \Lambda_\tau$ by $\hat{V}^\mu$ in the peeling arguments to obtain as before that

$$\mathbb{P}\left(M_\tau \geq \sqrt{2(1+\varepsilon)\hat{V}^\mu x} + x/3 \text{ and } w \leq \hat{V}^\mu \leq v \text{ and } \sup_{s\leq\tau,m} |H_s^{(m)}| \leq 1\right) \leq 2Ke^{-x}. \tag{7.12}$$

7.4. Proofs of the probabilistic results for Hawkes processes

7.4.1. Proof of Lemma 1

Let $K(n)$ denote the vector of the number of descendants in the n'th generation from a single ancestral point of type ℓ, define $K(0) = \mathbf{e}_\ell$ and let $W(n) = \sum_{k=0}^n K(k)$ denote the total number of points in the first n generations. Define for $\theta \in \mathbb{R}^M$

$$\phi_\ell(\theta) = \log \mathbb{E}_\ell e^{\theta^T K(1)}.$$

Thus, $\phi_\ell(\theta)$ is the log-Laplace transform of the distribution of $K(1)$ given that there is a single initial ancestral point of type ℓ. We define the vector $\phi(\theta)$ by $\phi(\theta)' = (\phi_1(\theta), ..., \phi_M(\theta))$. Note that ϕ only depends on the law of the number of children per parent, i.e. it only depends on Γ. Then

$$\begin{aligned}
\mathbb{E}_\ell e^{\theta^T W(n)} &= \mathbb{E}_\ell\left(e^{\theta^T W(n-1)}\mathbb{E}\left(e^{\theta^T K(n)} \mid K(n-1), \ldots, .K(1)\right)\right) \\
&= \mathbb{E}_\ell\left(e^{\theta^T W(n-1)} e^{\phi(\theta)^T K(n-1)}\right) \\
&= \mathbb{E}_\ell e^{(\theta+\phi(\theta))^T K(n-1)+\theta^T W(n-2)}
\end{aligned}$$

Defining $g(\theta) = \theta + \phi(\theta)$ we arrive by recursion at the formula

$$\begin{aligned}
\mathbb{E}_\ell e^{\theta^T W(n)} &= \mathbb{E}_\ell e^{g^{\circ(n-1)}(\theta)^T K(1)+\theta^T W(0)} \\
&= e^{\phi(g^{\circ(n-1)}(\theta))_\ell+\theta_\ell} \\
&= e^{g^{\circ n}(\theta)_\ell}.
\end{aligned}$$

where for any n, $g^{\circ(n)} = g \circ \cdots \circ g$ n times. Or, in other words, we have the following representation

$$\log \mathbb{E}_\ell e^{\theta^T W(n)} = g^{\circ n}(\theta)_\ell$$

of the log-Laplace transform of $W(n)$.

Below we show that ϕ is a contraction in a neighborhood containing 0, that is, for some $r > 0$ and a constant $C < 1$ (and a suitable norm), $||\phi(s)|| \leq C||s||$ for $||s|| \leq r$. If θ is chosen such that

$$\frac{||\theta||}{1-C} \leq r$$

we have $||\theta|| \leq r$, and if we assume that $g^{\circ k}(\theta) \in B(0,r)$ for $k = 1, \dots, n-1$ then

$$\begin{aligned} ||g^{\circ n}(\theta)|| &\leq ||\theta|| + ||\phi(g^{\circ(n-1)}(\theta))|| \\ &\leq ||\theta|| + C||g^{\circ(n-1)}(\theta)|| \\ &\leq ||\theta|| \left(1 + C + C^2 + \ldots + C^n\right) \\ &\leq r \end{aligned}$$

Thus, by induction, $g^{\circ n}(\theta) \in B(0,r)$ for all $n \geq 1$. Since $n \mapsto W_m(n)$ is increasing and goes to $W_m(\infty)$ for $n \to \infty$, with $W_m(\infty)$ the total number of points in a cluster of type m, and since $W = \sum_m W_m(\infty) = \mathbf{1}^T W(\infty)$, we have by monotone convergence that for $\vartheta \in \mathbb{R}$

$$\log \mathbb{E}_\ell e^{\vartheta W} = \lim_{n \to \infty} g^{\circ n}(\vartheta \mathbf{1})_\ell.$$

By the previous result, the right hand side is bounded if $|\vartheta|$ is sufficiently small. This completes the proof up to proving that ϕ is a contraction.

To this end we note that ϕ is continuously differentiable (on $\mathbb{R}^M$ in fact, but a neighborhood around 0 suffices) with derivative $D\phi(0) = \Gamma$ at 0. Since the spectral radius of Γ is strictly less than 1 there is a $C < 1$ and, by the Householder theorem, a norm $|| \cdot ||$ on $\mathbb{R}^M$ such that for the induced operator norm of Γ we have

$$||\Gamma|| = \max_{x:||x|| \leq 1} ||\Gamma x|| < C$$

Since the norm is continuous and $D\phi(s)$ is likewise there is an $r > 0$ such that

$$||D\phi(s)|| \leq C < 1$$

for $||s|| \leq r$. This, in turn, implies that ϕ is Lipschitz continuous in the ball $B(0,r)$ with Lipschitz constant C, and since $\phi(0) = 0$ we get

$$||\phi(s)|| \leq C||s||$$

for $||s|| \leq r$. This ends the proof of the lemma.

Note that we have not at all used the explicit formula for ϕ above, which is obtainable and simple since the offspring distributions are Poisson. The only thing we needed was the fact that ϕ is defined in a neighborhood around 0, thus that the offspring distributions are sufficiently light-tailed.

7.4.2. Proof of Proposition 2

We use the cluster representation, and we note that any cluster with ancestral point in $[-n-1,-n]$ must have at least $n+1-\lceil A\rceil$ points in the cluster if any of the points are to fall in $[-A,0)$. This follows from the assumption that all the $h_\ell^{(m)}$-functions have support in $[0,1]$. With $\tilde{N}_{A,\ell}$ the number of points in $[-A,0)$ from a cluster with ancestral points of type ℓ we thus have the bound

$$\tilde{N}_{A,\ell} \leq \sum_n \sum_{k=1}^{A_n} \max\{W_{n,k} - n + \lceil A\rceil, 0\}$$

where A_n is the number of ancestral points in $[-n-1,-n]$ of type ℓ and $W_{n,k}$ is the number of points in the respective clusters. Here the A_n's and the $W_{n,k}$'s are all independent, the A_n's are Poisson distributed with mean ν_ℓ and the $W_{n,k}$'s are i.i.d. with the same distribution as W in Lemma 1. Moreover,

$$H_n(\vartheta_\ell) := \mathbb{E}_\ell e^{\vartheta_\ell \max\{W-n+\lceil A\rceil,0\}} \leq \mathbb{P}_\ell(W \leq n - \lceil A\rceil) + e^{-\vartheta_\ell(n-\lceil A\rceil)}\mathbb{E}_\ell e^{\vartheta_\ell W},$$

which is finite for $|\vartheta_\ell|$ sufficiently small according to Lemma 1. Then we can compute an upper bound on the Laplace transform of $\tilde{N}_{A,\ell}$:

$$\begin{aligned}
\mathbb{E}e^{\vartheta_\ell \tilde{N}_{A,\ell}} &\leq \prod_n \mathbb{E} \prod_{k=1}^{A_n} \mathbb{E}\left(e^{\vartheta_\ell \max\{W_{n,k}-n+\lceil A\rceil,0\}} \mid A_n\right) \\
&\leq \prod_n \mathbb{E} H_n(\vartheta_\ell)^{A_n} \\
&= \prod_n e^{\nu_\ell(H_n(\vartheta_\ell)-1)} \\
&= e^{\nu_\ell \sum_n (H_n(\vartheta_\ell)-1)}
\end{aligned}$$

Since $H_n(\vartheta_\ell) - 1 \leq e^{-\vartheta_\ell(n-\lceil A\rceil)}\mathbb{E}_\ell e^{\vartheta_\ell W}$ we have $\sum_n (H_n(\vartheta_\ell)-1) < \infty$, which shows that the upper bound is finite. To complete the proof, observe that $N_{[-A,0)} = \sum_\ell \tilde{N}_{A,\ell}$ where $\tilde{N}_{A,\ell}$ for $\ell = 1,\ldots,M$ are independent. Since all variables are positive, it is sufficient to take $\theta = \min_\ell \vartheta_\ell$.

7.4.3. Proof of Proposition 3

In this paragraph, the notation $\square$ simply denotes a generic positive absolute constant that may change from line to line. The notation $\square_{\theta_1,\theta_2,\ldots}$ denotes a positive constant depending on $\theta_1, \theta_2, \ldots$ that may change from line to line.

Let

$$u = C_1 \sigma \log^{3/2}(T)\sqrt{T} + C_2 b(\log(T))^{2+\eta}, \tag{7.13}$$

where the choices of C_1 and C_2 will be given later. For any positive integer k such that $x := T/(2k) > A$, we have by stationarity:

$$\begin{aligned}\mathbb{P}\left(\int_0^T [Z \circ \mathfrak{S}_t(N) - \mathbb{E}(Z)]\mathrm{d}t \geq u\right) &= \mathbb{P}\left(\sum_{q=0}^{k-1} \int_{2qx}^{2qx+x} [Z \circ \mathfrak{S}_t(N) - \mathbb{E}(Z)]\mathrm{d}t \right.\\ &\quad \left. + \int_{2qx+x}^{2qx+2x} [Z \circ \mathfrak{S}_t(N) - \mathbb{E}(Z)]\mathrm{d}t \geq u\right)\\ &\leq 2\mathbb{P}\left(\sum_{q=0}^{k-1} \int_{2qx}^{2qx+x} [Z \circ \mathfrak{S}_t(N) - \mathbb{E}(Z)]\mathrm{d}t \geq \frac{u}{2}\right).\end{aligned}$$

Similarly to [51], we introduce $(\tilde{M}_q^x)_q$ a sequence of independent Hawkes processes, each being stationary with intensities per mark given by $\psi_t^{(m)}$. For each q, we then introduce M_q^x the truncated process associated with $\tilde{M}_q^x$, where truncation means that we only consider the points lying in $[2qx - A, 2qx + x]$. So, if we set

$$F_q = \int_{2qx}^{2qx+x} [Z \circ \mathfrak{S}_t(M_q^x) - \mathbb{E}(Z)]\mathrm{d}t,$$

$$\mathbb{P}\left(\int_0^T [Z \circ \mathfrak{S}_t(N) - \mathbb{E}(Z)]\mathrm{d}t \geq u\right) \leq 2\mathbb{P}\left(\sum_{q=0}^{k-1} F_q \geq \frac{u}{2}\right) + 2k\mathbb{P}\left(T_e > \frac{T}{2k} - A\right), \qquad (7.14)$$

where T_e represents the time to extinction of the process. More precisely T_e is the last point of the process if in the cluster representation only ancestral points before 0 are appearing. For more details, see section 3 of [51]. So, denoting a_l the ancestral points with marks l and $H_{a_l}^l$ the length of the corresponding cluster whose origin is a_l, we have:

$$T_e = \max_{l \in \{1,\dots,M\}} \max_{a_l} \left\{a_l + H_{a_l}^l\right\}.$$

But, for any $a > 0$,

$$\begin{aligned}\mathbb{P}(T_e \leq a) &= \mathbb{E}\left[\prod_{l=1}^M \prod_{a_l} \mathbb{E}\left[1_{\{a_l + H_{a_l}^l \leq a\}} | a_l\right]\right]\\ &= \mathbb{E}\left[\prod_{l=1}^M \prod_{a_l} \exp\left(\log\left(\mathbb{P}(H_0^l \leq a - a_l)\right)\right)\right]\\ &= \mathbb{E}\left[\prod_{l=1}^M \exp\left(\int_{-\infty}^0 \log(\mathbb{P}(H_0^l \leq a - x))\mathrm{d}\tilde{N}_x^{(l)}\right)\right],\end{aligned}$$

where $\tilde{N}^{(l)}$ denotes the process associated with the ancestral points with marks l. So,

$$\begin{aligned}\mathbb{P}(T_e \le a) &= \exp\left(\sum_{l=1}^{M}\int_{-\infty}^{0}\left(\exp(\log(\mathbb{P}(H_0^l \le a-x)))-1\right)\nu^{(l)}\mathrm{d}x\right)\\ &= \exp\left(-\sum_{l=1}^{M}\nu^{(l)}\int_{a}^{+\infty}\mathbb{P}(H_0^l > u)\mathrm{d}u\right).\end{aligned}$$

Now, by Lemma 1, there exists some $\vartheta_l > 0$, such that $c_l = \mathbb{E}_\ell(e^{\vartheta_l W}) < +\infty$, where W is the number of points in the cluster. But if all the interaction functions have support in $[0,1]$, one always have that $H_0^l < W$. Hence

$$\begin{aligned}\mathbb{P}(H_0^l > u) &\le \mathbb{E}[\exp(\vartheta_l H_0^l)]\exp(-\vartheta_l u)\\ &\le c_l\exp(-\vartheta_l u).\end{aligned}$$

So,

$$\begin{aligned}\mathbb{P}(T_e \le a) &\ge \exp\left(-\sum_{l=1}^{M}\nu^{(l)}\int_{a}^{+\infty}c_l\exp(-\vartheta_l u)\mathrm{d}u\right)\\ &= \exp\left(-\sum_{l=1}^{M}\nu^{(l)}c_l/\vartheta_l\exp(-\vartheta_l a)\right)\\ &\ge 1-\sum_{l=1}^{M}\nu^{(l)}c_l/\vartheta_l\exp(-\vartheta_l a).\end{aligned}$$

So, there exists a constant $C_{\alpha,f_0,A}$ depending on α, A, and f_0 such that if we take $k = \lfloor C_{\alpha,A,f_0}T/\log(T)\rfloor$, then

$$k\mathbb{P}\left(T_e > \frac{T}{2k}-A\right) \le T^{-\alpha}.$$

In this case $x = \frac{T}{2k} \approx \log(T)$ is larger than A for T large enough (depending on A, α, f_0).

Now, let us focus on the first term B of (7.14), where

$$B = \mathbb{P}\left(\sum_{q=0}^{k-1}F_q \ge \frac{u}{2}\right).$$

Let us consider some $\tilde{\mathcal{N}}$ where $\tilde{\mathcal{N}}$ will be fixed later and let us define the measurable events

$$\Omega_q = \left\{\sup_t\{M_q^x|_{[t-A,t)}\} \le \tilde{\mathcal{N}}\right\},$$

where $M_q^x|_{[t-A,t)}$ represents the set of points of M_q^x lying in $[t-A,t)$. Let us also consider $\Omega = \cap_{1\le q\le k}\Omega_q$. Then

$$B \le \mathbb{P}\big(\sum_q F_q \ge u/2 \text{ and } \Omega\big) + \mathbb{P}(\Omega^c).$$

We have $\mathbb{P}(\Omega^c) \leq \sum_q \mathbb{P}(\Omega_q^c)$. Each Ω_q can also be easily controlled. Indeed it is sufficient to split $[2qx - A, 2qx + x]$ in intervals of size A (there are about $\square_{\alpha,A,f_0} \log(T)$ of those) and require that the number of points in each subinterval is smaller than $\tilde{\mathcal{N}}/2$. By stationarity, we obtain that

$$\mathbb{P}(\Omega_q^c) \leq \square_{\alpha,A,f_0} \log(T) \mathbb{P}(N_{[-A,0)} > \tilde{\mathcal{N}}/2).$$

Using Proposition 2 with $u = \lceil \tilde{\mathcal{N}}/2 \rceil + 1/2$, we obtain:

$$\mathbb{P}(\Omega_q^c) \leq \square_{\alpha,A,f_0} \log(T) \exp(-\square_{\alpha,A,f_0} \tilde{\mathcal{N}}) \text{ and } \mathbb{P}(\Omega^c) \leq \square_{\alpha,A,f_0} T \exp(-\square_{\alpha,A,f_0} \tilde{\mathcal{N}}). \tag{7.15}$$

Note that this control holds for any positive choice of $\tilde{\mathcal{N}}$. Hence this gives also the following Lemma that will be used later.

Lemma 3. *For any $\mathcal{R} > 0$,*

$$\mathbb{P}\big(\textit{there exists } t \in [0,T] \mid \quad M_q^x|_{[t-A,t)} > \mathcal{R}) \leq \square_{\alpha,A,f_0} T \exp(-\square_{\alpha,A,f_0} \mathcal{R}).$$

Hence by taking $\tilde{\mathcal{N}} = C_3 \log(T)$ for C_3 large enough this is smaller than $\square_{\alpha,A,f_0} T^{-\alpha'}$, where $\alpha' = \max(\alpha, 2)$.

It remains to obtain the rate of $D := \mathbb{P}(\sum_q F_q \geq u/2 \text{ and } \Omega)$. For any positive constant θ that will be chosen later, we have:

$$\begin{aligned} D &\leq e^{-\frac{\theta u}{2}} \mathbb{E}\left(e^{\theta \sum_q F_q} \prod_q \mathbb{1}_{\Omega_q} \right) \\ &\leq e^{-\frac{\theta u}{2}} \prod_q \mathbb{E}\left(e^{\theta F_q} \mathbb{1}_{\Omega_q} \right) \end{aligned} \tag{7.16}$$

since the variables $(M_q^x)_q$ are independent. But

$$\mathbb{E}\left(e^{\theta F_q} \mathbb{1}_{\Omega_q} \right) = 1 + \theta \mathbb{E}(F_q \mathbb{1}_{\Omega_q}) + \sum_{j \geq 2} \frac{\theta^j}{j!} \mathbb{E}(F_q^j \mathbb{1}_{\Omega_q})$$

and $\mathbb{E}(F_q \mathbb{1}_{\Omega_q}) = \mathbb{E}(F_q) - \mathbb{E}(F_q \mathbb{1}_{\Omega_q^c}) = -\mathbb{E}(F_q \mathbb{1}_{\Omega_q^c})$.

Next note that if for any integer l,

$$l\tilde{\mathcal{N}} < \sup_t M_q^x|_{[t-A,t)} \leq (l+1)\tilde{\mathcal{N}}$$

then

$$|F_q| \leq xb[(l+1)^\eta \tilde{\mathcal{N}}^\eta + 1] + x\mathbb{E}(Z).$$

Hence, cutting Ω_q^c in slices of the type $\{l\tilde{\mathcal{N}} < \sup_t M_q^x{}_{[t-A,t)} \leq (l+1)\tilde{\mathcal{N}}\}$ and using Lemma 3, we obtain by taking C_3 large enough,

$$\begin{aligned}
|\mathbb{E}(F_q \mathbb{1}_{\Omega_q})| = |\mathbb{E}(F_q \mathbb{1}_{\Omega_q^c})| &\leq \sum_{l=1}^{+\infty} x(b[(l+1)^\eta \tilde{\mathcal{N}}^\eta + 1] + |\mathbb{E}(Z)|) \times \\
&\qquad \mathbb{P}(\text{there exists } t \in [0,T] \mid \{M_q^x|_{[t-A,t)}\} > \ell\tilde{\mathcal{N}}) \\
&\leq \square_{\alpha,A,f_0} \sum_{l=1}^{+\infty} x(b[(l+1)^\eta \tilde{\mathcal{N}}^\eta + 1] + |\mathbb{E}(Z)|) \log(T) e^{-\square_{\alpha,A,f_0} l\tilde{\mathcal{N}}} \\
&\leq \square_{\alpha,A,f_0} \sum_{l=1}^{+\infty} x(b\tilde{\mathcal{N}}^\eta + |\mathbb{E}(Z)|) \log(T) 2^{l\eta} e^{-\square_{\alpha,A,f_0} l\tilde{\mathcal{N}}} \\
&\leq \square_{\alpha,\eta,A,f_0} \log^2(T) b\tilde{\mathcal{N}}^\eta \frac{e^{-\square_{\alpha,A,f_0}\tilde{\mathcal{N}}}}{1 - 2^\eta e^{-\square_{\alpha,A,f_0}\tilde{\mathcal{N}}}} \\
&\leq z_1 := \square_{\alpha,\eta,A,f_0} b T^{-\alpha'}.
\end{aligned}$$

Note that in the previous inequalities, we have bounded $|\mathbb{E}(Z)|$ by $b\mathbb{E}[N_{[-A,0)}^\eta]$. In the same way, one can bound

$$\mathbb{E}(F_q^j \mathbb{1}_{\Omega_q}) \leq \mathbb{E}(F_q^2 \mathbb{1}_{\Omega_q}) z_b^{j-2},$$

with $z_b := xb[\tilde{\mathcal{N}}^\eta + 1] + x\mathbb{E}(Z) = \square_{\alpha,\eta,A,f_0} b \log(T)^{1+\eta}$. One can also note that by stationarity,

$$\begin{aligned}
\mathbb{E}(F_q^2 \mathbb{1}_{\Omega_q}) &\leq x\mathbb{E}\left[\int_{2qx}^{2qx+x} [Z \circ \theta_s(M_q^x) - \mathbb{E}(Z)]^2 \mathbb{1}_{\{\text{for all } t, M_q^x|_{[t-A,t)} \leq \tilde{\mathcal{N}}\}} \mathrm{d}s\right] \\
&\leq x\mathbb{E}\left[\int_{2qx}^{2qx+x} [Z \circ \theta_s(M_q^x) - \mathbb{E}(Z)]^2 \mathbb{1}_{\{M_q^x|_{[s-A,s)} \leq \tilde{\mathcal{N}}\}} \mathrm{d}s\right] \\
&\leq x^2 \mathbb{E}([Z(N) - \mathbb{E}(Z)]^2 \mathbb{1}_{N_{[-A,0)} \leq \tilde{\mathcal{N}}}) \\
&\leq z_v := \square_{\alpha,\eta,A,f_0} (\log(T))^2 \sigma^2.
\end{aligned}$$

Now let us go back to (7.16). We have that

$$\begin{aligned}
D &\leq \exp\left[-\frac{\theta u}{2} + k \ln\left(1 + \theta z_1 + \sum_{j \geq 2} z_v z_b^{j-2} \frac{\theta^j}{j!}\right)\right] \\
&\leq \exp\left[-\theta\left(\frac{u}{2} - kz_1\right) + k \sum_{j \geq 2} z_v z_b^{j-2} \frac{\theta^j}{j!}\right],
\end{aligned}$$

using that $\ln(1+u) \leq u$. It is sufficient now to recognize a step of the proof of the Bernstein inequality (weak version see [41, p25]). Since $kz_1 = \square_{\alpha,\eta,s} bT^{1-\alpha'}/(\log(T))$, one can choose $\alpha' > 1, C_1$ and C_2 in the definition (7.13) of u (not depending on b) such that $u/2 - kz_1 \geq \sqrt{2kz_v z} + \frac{1}{3} z_b z$ for some $z = C_4 \log(T)$, where C_4 is a constant. Hence

$$D \leq \exp\left[-\theta(\sqrt{2kz_v z} + \frac{1}{3} z_b z) + k \sum_{j\geq 2} z_v z_b^{j-2} \frac{\theta^j}{j!}\right].$$

One can choose accordingly θ (as for the proof of the Bernstein inequality) to obtain a bound in e^{-z}. It remains to choose C_4 large enough and only depending on α, η, A and f_0 to guarantee that $D \leq e^{-z} \leq \square_{\alpha,\eta,A,f_0} T^{-\alpha}$. This concludes the proof of the proposition.

7.4.4. Proof of Proposition 4

Let $\mathbb{Q}$ denote a measure such that under $\mathbb{Q}$ the distribution of the full point process restricted to $(-\infty, 0]$ is identical to the distribution under $\mathbb{P}$ and such that on $(0,\infty)$ the process consists of independent components each being a homogeneous Poisson process with rate 1. Furthermore, the Poisson processes should be independent of the process on $(-\infty, 0]$. From Corollary 5.1.2 in [36] the likelihood process is given by

$$\mathcal{L}_t = \exp\left(Mt - \sum_m \int_0^t \lambda_u^{(m)} \mathrm{d}u + \sum_m \int_0^t \log \lambda_u^{(m)} \mathrm{d}N_u^{(m)}\right)$$

and we have for $t \geq 0$ the relation

$$\mathbb{E}_{\mathbb{P}} \kappa_t(\mathbf{f})^2 = \mathbb{E}_{\mathbb{Q}} \kappa_t(\mathbf{f})^2 \mathcal{L}_t, \tag{7.17}$$

where $\mathbb{E}_{\mathbb{P}}$ and $\mathbb{E}_{\mathbb{Q}}$ denote the expectation with respect to $\mathbb{P}$ and $\mathbb{Q}$ respectively. Let, furthermore, $\tilde{N}_1 = N_{[-1,0)}$ denote the total number of points on $[-1, 0)$. Proposition 4 will be an easy consequence of the following lemma.

Lemma 4. *If the point process is stationary under $\mathbb{P}$, if*

$$e^d \leq \lambda_t^{(m)} \leq a(N_1 + \tilde{N}_1) + b$$

for $t \in [0,1]$ and for constants $d \in \mathbb{R}$ and $a, b > 0$, and if $\mathbb{E}_{\mathbb{P}}(1+\varepsilon)^{\tilde{N}_1} < \infty$ for some $\varepsilon > 0$ then for any $\mathbf{f}$,

$$Q(\mathbf{f},\mathbf{f}) \geq \zeta \|\mathbf{f}\|^{\mathbf{2}} \tag{7.18}$$

for some constant $\zeta > 0$.

Proof. We use Hölders inequality on $\kappa_1(\mathbf{f})^{\frac{2}{p}} \mathcal{L}_1^{\frac{1}{p}}$ and $\kappa_1(\mathbf{f})^{\frac{2}{q}} \mathcal{L}_1^{-\frac{1}{p}}$ to get

$$\mathbb{E}_{\mathbb{Q}} \kappa_1(\mathbf{f})^2 \leq \left(\mathbb{E}_{\mathbb{Q}} \kappa_1(\mathbf{f})^2 \mathcal{L}_1\right)^{\frac{1}{p}} \left(\mathbb{E}_{\mathbb{Q}} \kappa_1(\mathbf{f})^2 \mathcal{L}_1^{-\frac{q}{p}}\right)^{\frac{1}{q}} = Q(\mathbf{f},\mathbf{f})^{\frac{1}{p}} \left(\mathbb{E}_{\mathbb{Q}} \kappa_1(\mathbf{f})^2 \mathcal{L}_1^{1-q}\right)^{\frac{1}{q}} \tag{7.19}$$

where $\frac{1}{p} + \frac{1}{q} = 1$. We choose $q \geq 1$ (and thus p) below to make $q - 1$ sufficiently small. For the left hand side we have by independence of the homogeneous Poisson processes that if $\mathbf{f} = (\mu, (g_\ell)_{\ell=1,\dots,M})$,

$$\begin{aligned}
\mathbb{E}_{\mathbb{Q}}\kappa_1(\mathbf{f})^2 &= (\mathbb{E}_{\mathbb{Q}}\kappa_1(\mathbf{f}))^2 + \mathbb{V}_{\mathbb{Q}}\kappa_1(\mathbf{f}) \\
&= \left(\mu + \sum_\ell \int_0^1 g_\ell(u)\mathrm{d}u\right)^2 + \sum_\ell \int_0^1 g_\ell(u)^2 \mathrm{d}u.
\end{aligned}$$

Exactly as on page 32 in [52] there exists $c' > 0$ such that

$$\mathbb{E}_{\mathbb{Q}}\kappa_1(\mathbf{f})^2 \geq c' \left(\mu^2 + \sum_\ell \int_0^1 g_\ell^2(u)\mathrm{d}u\right) = c'||\mathbf{f}||^2. \tag{7.20}$$

To bound the second factor on the right hand side in (7.19) we observe, by assumption, that we have the lower bound

$$\mathcal{L}_1 \geq e^{M(1-b)} e^{(d-aM)N_1} e^{-aM\tilde{N}_1}$$

on the likelihood process. Under $\mathbb{Q}$ we have that $(\kappa_1(\mathbf{f}), N_1)$ and $\tilde{N}_1$ are independent, and with $\rho = e^{(q-1)(aM-d)}$ and $\tilde{\rho} = e^{(q-1)(aM)}$ we get that

$$\mathbb{E}_{\mathbb{Q}}\kappa_1(\mathbf{f})^2 \mathcal{L}_1^{1-q} \leq e^{(q-1)M(b-1)} \mathbb{E}_{\mathbb{Q}}\tilde{\rho}^{\tilde{N}_1} \mathbb{E}_{\mathbb{Q}}\kappa_1(\mathbf{f})^2 \rho^{N_1}.$$

Here we choose q such that $\tilde{\rho}$ is sufficiently close to 1 to make sure that $\mathbb{E}_{\mathbb{Q}}\tilde{\rho}^{\tilde{N}_1} = \mathbb{E}_{\mathbb{P}}\tilde{\rho}^{\tilde{N}_1} < \infty$ (see Proposition 2). Moreover, by Cauchy-Schwarz' inequality

$$\kappa_1^2(\mathbf{f}) \leq \left(\mu^2 + \sum_\ell \int_0^{1-} g_\ell^2(1-u)\mathrm{d}N_u^{(\ell)}\right)(1 + N_1). \tag{7.21}$$

Under $\mathbb{Q}$ the point processes on $(0, \infty)$ are homogeneous Poisson processes with rate 1 and N_1, the total number of points, is Poisson. This implies that conditionally on $(N_1^{(1)}, \dots, N_1^{(M)}) = (n^{(1)}, \dots, n^{(M)})$ the $n^{(m)}$-points for the m'th process are uniformly distributed on $[0, 1]$, hence

$$\mathbb{E}_{\mathbb{Q}}\kappa_1(\mathbf{f})^2 \mathcal{L}_1^{1-q} \leq \left(\mu^2 + \sum_\ell \int_0^1 g_\ell^2(u)\mathrm{d}u\right) \underbrace{e^{(q-1)M(b-1)} \mathbb{E}_{\mathbb{Q}}\tilde{\rho}^{\tilde{N}_1} \mathbb{E}_{\mathbb{Q}}(1+N_1)^2 \rho^{N_1}}_{c''} = c''||\mathbf{f}||^2. \tag{7.22}$$

Combining (7.20) and (7.22) with (7.19) we get that

$$c'||\mathbf{f}||^2 \leq (c'')^{\frac{1}{q}} ||\mathbf{f}||^{\frac{2}{q}} Q(\mathbf{f}, \mathbf{f})^{\frac{1}{p}}$$

or by rearranging that

$$Q(\mathbf{f}, \mathbf{f}) \geq \zeta ||\mathbf{f}||^2$$

with $\zeta = (c')^p / (c'')^{p-1}$. □

For the Hawkes process it follows that if $\nu^{(m)} > 0$ and if

$$\sup_{t\in[0,1]} h_\ell^{(m)}(t) < \infty$$

for $l, m = 1, \ldots, M$ then for $t \in [0,1]$ we have $e^d \leq \lambda_t^{(m)} \leq a(N_1 + \tilde{N}_1) + b$ with

$$d = \log \nu^{(m)}, \quad a = \max_l \sup_{t\in[0,1]} h_\ell^{(m)}(t), \quad b = \nu^{(m)}.$$

Proposition 2 proves that there exists $\varepsilon > 0$ such that $\mathbb{E}_{\mathbb{P}}(1+\varepsilon)^{\tilde{N}_1} < \infty$. This completes the proof of Proposition 4.

7.5. Proofs of the results of Sections 4.2 and 5.2

7.5.1. Proof of Propositions 5 and 1

We first prove Proposition 5. As in the proof of Proposition 3, we use the notation $\square$. Note that for any φ_1 and any φ_2 belonging to Φ,

$$G_{\varphi_1,\varphi_2} = \sum_{m=1}^{M} \int_0^T \kappa_t({\boldsymbol{\varphi_1}}^{(m)})\kappa_t({\boldsymbol{\varphi_2}}^{(m)})\mathrm{d}t$$

and $\mathbb{E}(G_{\varphi_1,\varphi_2}) = T\sum_{m=1}^{M} Q({\boldsymbol{\varphi_1}}^{(m)}, {\boldsymbol{\varphi_2}}^{(m)})$ by using (5.2). This implies that

$$\mathbb{E}(a'Ga) = a'\mathbb{E}(G)a = T\sum_m Q(\mathbf{f}_a^{(m)}, \mathbf{f}_a^{(m)}).$$

Hence by Proposition 4, $\mathbb{E}(a'Ga) \geq T\zeta\sum_m \|\mathbf{f}_a^{(m)}\|^2 = T\zeta\|f_a\|^2$ by definition of the norm on $\mathcal{H}$. Since Φ is an orthonormal system, this implies that $\mathbb{E}(a'Ga) \geq T\zeta\|a\|_{\ell_2}$. Hence, to show that Ω_c is a large event for some $c > 0$, it is sufficient to show that for some $0 < \epsilon < \zeta$, with high probability, for any $a \in \mathbb{R}^\Phi$,

$$|a'Ga - a'\mathbb{E}(G)a| \leq T\epsilon\|a\|_{\ell_2}^2. \tag{7.23}$$

Indeed, (7.23) implies that, with high probability, for any $a \in \mathbb{R}^\Phi$,

$$a'Ga \geq a'\mathbb{E}(G)a - T\epsilon\|a\|_{\ell_2} \geq T(\zeta - \epsilon)\|a\|_{\ell_2},$$

and the choice $c = T(\zeta - \epsilon)$ is convenient. So, first one has to control all the coefficients of $G - \mathbb{E}(G)$. For all $\varphi, \rho \in \Phi$, we apply Proposition 3 to

$$Z(N) = \sum_m \psi_0^{(m)}(\varphi)\psi_0^{(m)}(\rho).$$

Note that Z only depends on points lying in $[-1,0)$. Therefore, $|Z(N)| \leq 2M\|\varphi\|_\infty\|\rho\|_\infty\big(1+N^2_{[-1,0)}\big)$. This leads to

$$\mathbb{P}\left(\frac{1}{T}\left|G_{\varphi,\rho} - \mathbb{E}(G_{\varphi,\rho})\right| \geq x_{\varphi,\rho}\right) \leq \square_{\beta,f_0} T^{-\beta}$$

with

$$x_{\varphi,\rho} = \square_{\beta,f_0,M}[\sigma_{\varphi,\rho}\log^{3/2}(T)T^{-1/2} + \|\varphi\|_\infty\|\rho\|_\infty \log^4(T)T^{-1}]$$

and

$$\sigma^2_{\varphi,\rho} = \mathbb{E}\left[\left[\sum_m \psi_0^{(m)}(\varphi)\psi_0^{(m)}(\rho) - \mathbb{E}\Big(\sum_m \psi_0^{(m)}(\varphi)\psi_0^{(m)}(\rho)\Big)\right]^2 \mathbb{1}_{N_{[-1,0)}\leq\tilde{\mathcal{N}}}\right].$$

Hence, with probability larger than $1-\square_{\beta,f_0}|\Phi|^2T^{-\beta}$ one has that

$$|a'Ga - a'\mathbb{E}(G)a| \leq \square_{\beta,f_0}\left(\sum_{\varphi,\rho\in\Phi}|a_\varphi||a_\rho|[\sigma_{\varphi,\rho}\log^{3/2}(T)T^{1/2} + \|\varphi\|_\infty\|\rho\|_\infty\log^4(T)]\right).$$

Hence, for any positive constant δ chosen later,

$$|a'Ga - a'\mathbb{E}(G)a| \leq \square_{\beta,f_0}\left[T\sum_{\varphi,\rho\in\Phi}|a_\varphi||a_\rho|\left[\delta\frac{\sigma^2_{\varphi,\rho}}{\|\varphi\|_\infty\|\rho\|_\infty} + \left[\frac{1}{\delta\log(T)}+1\right]\|\varphi\|_\infty\|\rho\|_\infty\frac{\log^4(T)}{T}\right]\right]. \quad (7.24)$$

Now let us focus on $E := \sum_{\varphi,\rho\in\Phi}|a_\varphi||a_\rho|\frac{\sigma^2_{\varphi,\rho}}{\|\varphi\|_\infty\|\rho\|_\infty}$. First, we have:

$$E \leq 2\sum_{\varphi,\rho\in\Phi}|a_\varphi||a_\rho|\frac{\mathbb{E}([\sum_m\psi_0^{(m)}(\varphi)\psi_0^{(m)}(\rho)]^2\mathbb{1}_{N_{[-1,0)}\leq\tilde{\mathcal{N}}}) + (\mathbb{E}[\sum_m\psi_0^{(m)}(\varphi)\psi_0^{(m)}(\rho)])^2}{\|\varphi\|_\infty\|\rho\|_\infty}$$

with $\tilde{\mathcal{N}} := \square_{\beta,f_0}\log(T)$. Next,

$$\sum_m \psi_0^{(m)}(\varphi)\psi_0^{(m)}(\rho) \leq 2M\|\varphi\|_\infty\|\rho\|_\infty(1+N^2_{[-1,0)}).$$

Hence, if $N_{[-1,0)} \leq \tilde{\mathcal{N}} = \square_{\beta,f_0}\log(T)$, for T large enough,

$$\sum_m \psi_0^{(m)}(\varphi)\psi_0^{(m)}(\rho) \leq \square_{\beta,M,f_0}\|\varphi\|_\infty\|\rho\|_\infty\log^2(T)$$

and

$$\mathbb{E}(\sum_m \psi_0^{(m)}(\varphi)\psi_0^{(m)}(\rho)) \leq \square_{\beta,M,f_0}\|\varphi\|_\infty\|\rho\|_\infty\log^2(T).$$

Hence,

$$E \leq \square_{\beta,M,f_0} \log^2(T) \sum_{\varphi,\rho\in\Phi} |a_\varphi||a_\rho| \mathbb{E}\left(\Big| \sum_m \psi_0^{(m)}(\varphi)\psi_0^{(m)}(\rho) \Big| \right).$$

But note that for any f, $|\psi_0^{(m)}(f)| \leq \psi_0^{(m)}(|f|)$ where $|f| = ((|\mu^{(m)}|, (|g_\ell^{(m)}|)_{\ell=1,\dots,M})_{m=1,\dots,M}$. Therefore,

$$\begin{aligned} E &\leq \square_{\beta,M,f_0} \log^2(T) \sum_{\varphi,\rho\in\Phi} |a_\varphi||a_\rho| \mathbb{E}\left(\sum_m \psi_0^{(m)}(|\varphi|)\psi_0^{(m)}(|\rho|) \Big| \right) \\ &\leq \square_{\beta,M,f_0} \log^2(T) \sum_m \mathbb{E}\left(\left[\sum_{\varphi\in\Phi} |a_\varphi|\psi_0^{(m)}(|\varphi|) \right]^2 \right) \\ &\leq \square_{\beta,M,f_0} \log^2(T) \sum_m \mathbb{E}\left(\left[\psi_0^{(m)}\Big(\sum_{\varphi\in\Phi} |a_\varphi||\varphi| \Big) \right]^2 \right). \end{aligned}$$

But if $\varphi = (\mu_\varphi^{(m)}, ((g_\varphi)_\ell^{(m)})_\ell)_m$, then

$$\left[\psi_0^{(m)}\Big(\sum_{\varphi\in\Phi} |a_\varphi||\varphi| \Big) \right]^2 = \left[\sum_\varphi |a_\varphi|\mu_\varphi^{(m)} + \sum_{\ell=1}^M \int_{-1}^{0-} \sum_\varphi |a_\varphi||(g_\varphi)_\ell^{(m)}|(-u) \mathrm{d}N_u^{(\ell)} \right]^2.$$

If one creates artificially a process $N^{(0)}$ with only one point and if we decide that $(g_\varphi)_0^{(m)}$ is the constant function equal to $\mu_\varphi^{(m)}$, this can also be rewritten as

$$\left[\psi_0^{(m)}\Big(\sum_{\varphi\in\Phi} |a_\varphi||\varphi| \Big) \right]^2 = \left[\sum_{\ell=0}^M \int_{-1}^{0-} \sum_\varphi |a_\varphi||(g_\varphi)_\ell^{(m)}|(-u) \mathrm{d}N_u^{(\ell)} \right]^2.$$

Now we apply the Cauchy-Schwarz inequality for the measure $\sum_\ell \mathrm{d}N^{(\ell)}$, which gives

$$\left[\psi_0^{(m)}\Big(\sum_{\varphi\in\Phi} |a_\varphi||\varphi| \Big) \right]^2 \leq (N_{[-1,0)}+1) \sum_{\ell=0}^M \int_{-1}^{0-} \left[\sum_\varphi |a_\varphi||(g_\varphi)_\ell^{(m)}|(-u) \right]^2 \mathrm{d}N_u^{(\ell)}.$$

Consequently,

$$\begin{aligned} E &\leq \square_{\beta,M,f_0} \log^2(T) \sum_{m=1}^M \sum_{\ell=0}^M \mathbb{E}\left((N_{[-1,0)}+1) \int_{-1}^{0-} \left[\sum_\varphi |a_\varphi||(g_\varphi)_\ell^{(m)}|(-u) \right]^2 \mathrm{d}N_u^{(\ell)} \right) \\ &\leq \square_{\beta,M,f_0} \log^2(T) \sum_{m=1}^M \sum_{\ell=0}^M \sum_{\varphi,\rho\in\Phi} |a_\varphi||a_\rho| \times \\ &\qquad \mathbb{E}\left(\int_{-1}^{0-} (N_{[-1,0)}+1)|(g_\varphi)_\ell^{(m)}|(-u)|(g_\rho)_\ell^{(m)}|(-u) \mathrm{d}N_u^{(\ell)} \right). \end{aligned}$$

Now let us use the fact that for every $x, y \geq 0$, $\eta, \theta > 0$ that will be chosen later,

$$xy - \eta e^{\theta x} \leq \frac{y}{\theta}\left[\log(y) - \log(\eta\theta) - 1\right],$$

with the convention that $y\log(y) = 0$ if $y = 0$. Let us apply this to $x = N_{[-1,0)} + 1$ and $y = |(g_\varphi)_\ell^{(m)}|(-u)|(g_\rho)_\ell^{(m)}|(-u)$. We obtain that

$$E \leq \square_{\beta,M,f_0}\eta\log^2(T)\sum_{m=1}^{M}\sum_{\varphi,\rho\in\Phi}|a_\varphi||a_\rho|\mathbb{E}\left((N_{[-1,0)}+1)e^{\theta(N_{[-1,0)}+1)}\right)+$$

$$\square_{\beta,M,f_0}\theta^{-1}\log^2(T)\sum_{m=1}^{M}\sum_{\ell=0}^{M}\sum_{\varphi,\rho\in\Phi}|a_\varphi||a_\rho|\times$$

$$\mathbb{E}\left(\int_{-1}^{0^-}|(g_\varphi)_\ell^{(m)}||(g_\rho)_\ell^{(m)}|(-u)\left[\log(|(g_\varphi)_\ell^{(m)}||(g_\rho)_\ell^{(m)}|(-u)) - \log(\eta\theta) - 1\right]\mathrm{d}N_u^\ell\right).$$

Since for $\ell > 0$, $\mathrm{d}N_u^{(\ell)}$ is stationnary, one can replace $\mathbb{E}(\mathrm{d}N_u^{(\ell)})$ by $\square_{f_0}du$. Moreover since by Proposition 2, $N_{[-1,0)}$ has some exponential moments there exists $\theta = \square_{f_0}$ such that $\mathbb{E}\left((N_{[-1,0)}+1)e^{\theta(N_{[-1,0)}+1)}\right) = \square_{f_0}$. With $|\Phi|$ the size of the dictionary, this leads to

$$E \leq \square_{\beta,M,f_0}\eta|\Phi|\log^2(T)\|a\|_{\ell_2}^2+$$

$$\square_{\beta,M,f_0}\log^2(T)\sum_{m=1}^{M}\left[\sum_{\varphi,\rho\in\Phi}|a_\varphi||a_\rho||\mu_\varphi^{(m)}||\mu_\rho^{(m)}|\left[\log(|\mu_\varphi^{(m)}||\mu_\rho^{(m)}|) - \log(\eta\theta) - 1\right]+\right.$$

$$\left.\sum_{\ell=1}^{M}\sum_{\varphi,\rho\in\Phi}|a_\varphi||a_\rho|\int_0^1|(g_\varphi)_\ell^{(m)}||(g_\rho)_\ell^{(m)}|(u)\left[\log(|(g_\varphi)_\ell^{(m)}||(g_\rho)_\ell^{(m)}|(u)) - \log(\eta\theta) - 1\right]\mathrm{d}u\right].$$

Consequently, using $\|\Phi\|_\infty$ and r_Φ,

$$E \leq \square_{\beta,M,f_0}\eta|\Phi|\log^2(T)\|a\|_{\ell_2}^2 + \square_{\beta,M,f_0}\log^2(T)r_\Phi[2\log(\|\Phi\|_\infty) - \log(\eta\theta) - 1]\|a\|_{\ell_2}^2.$$

We choose $\eta = |\Phi|^{-1}$ and obtain that

$$E \leq \square_{\beta,M,f_0}\log^2(T)r_\Phi[\log(\|\Phi\|_\infty) + \log(|\Phi|)]\|a\|_{\ell_2}^2.$$

Now, let us choose $\delta = \omega/(\log^2(T)r_\Phi[\log(\|\Phi\|_\infty) + \log(|\Phi|)])$ where ω depends only on β, M and f_0 and will be chosen later and let us go back to (7.24):

$$\begin{aligned}\frac{1}{T}|a'Ga - a'\mathbb{E}(G)a| &\leq \square_{\beta,M,f_0}\omega\|a\|_{\ell_2}^2 + \square_{\beta,f_0,\omega}r_\Phi[\log(\|\Phi\|_\infty) \\ &\qquad + \log(|\Phi|)]\sum_{\varphi,\rho\in\Phi}|a_\varphi||a_\rho|\|\varphi\|_\infty\|\rho\|_\infty\frac{\log^5(T)}{T} \\ &\leq \square_{\beta,M,f_0}\omega\|a\|_{\ell_2}^2 + \square_{\beta,f_0,\omega}\|a\|_{\ell_2}^2A_\Phi(T).\end{aligned}$$

Under assumptions of Proposition 5, for T_0 large enough and $T \geq T_0$,

$$\frac{1}{T}|a'Ga - a'\mathbb{E}(G)a| \leq \square_{\beta,M,f_0}\omega\|a\|_{\ell_2}^2.$$

It is now sufficient to take ω small enough and then T_0 large enough to obtain (7.23) with $\epsilon < \zeta$ and Proposition 5 is proved.

Arguments for the proof of Proposition 1 are similar. So we just give a brief sketch of the proof. Now,

$$G_{\varphi_1,\varphi_2} = \sum_{m=1}^{M}\int_0^1 \left(Y_t^{(m)}\right)^2\varphi_1(t,X^{(m)})\varphi_2(t,X^{(m)})\mathrm{d}t.$$

Let $\beta > 0$. With probability larger than $1 - 2M^{-\beta}$,

$$\frac{1}{M}|G_{\varphi_1,\varphi_2} - \mathbb{E}[G_{\varphi_1,\varphi_2}]| \leq \sqrt{\frac{2\beta v_{\varphi_1,\varphi_2}\log M}{M}} + \frac{\beta b_{\varphi_1,\varphi_2}\log M}{3M},$$

with

$$b_{\varphi_1,\varphi_2} = \|\varphi_1\|_\infty\|\varphi_2\|_\infty,$$

$$v_{\varphi_1,\varphi_2} = \mathbb{E}\left(\int_0^1 \left(Y_t^{(m)}\right)^2\varphi_1(t,X^{(m)})\varphi_2(t,X^{(m)})\mathrm{d}t\right)^2 \leq D\|\varphi_1\|_\infty\|\varphi_2\|_\infty\langle|\varphi_1|,|\varphi_2|\rangle,$$

where $\langle\cdot,\cdot\rangle$ denotes the standard $\mathbb{L}_2$-scalar product. We have just used the classical Bernstein inequality combined with (4.2). So, with probability larger than $1 - 2|\Phi|^2M^{-\beta}$, for any vector a and any $\delta > 0$,

$$\begin{aligned}|a'Ga - \mathbb{E}[a'Ga]| &\leq \square_{D,\beta}\sum_{\varphi_1,\varphi_2}|a_{\varphi_1}||a_{\varphi_2}|[\delta M\langle|\varphi_1|,|\varphi_2|\rangle + \delta^{-1}\log M\|\varphi_1\|_\infty\|\varphi_2\|_\infty] \\ &\leq \square_{D,\beta}(\delta M r_\Phi + \delta^{-1}\|\Phi\|_\infty^2|\Phi|\log M)\|a\|_{\ell_2}^2.\end{aligned}$$

We choose $\delta = \sqrt{\frac{\|\Phi\|_\infty^2|\Phi|\log M}{Mr_\Phi}}$, so that with probability larger than $1 - 2|\Phi|^2M^{-\beta}$,

$$\frac{1}{M}|a'Ga - \mathbb{E}[a'Ga]| \leq \square_{D,\beta}\sqrt{\frac{\|\Phi\|_\infty^2 r_\Phi|\Phi|\log M}{M}}\|a\|_{\ell_2}^2.$$

We use (4.1) and (4.3) to conclude as for Proposition 5 and we obtain Proposition 1.

7.5.2. Proof of Corollary 3

First let us cut $[-1,T]$ in $\lfloor T\rfloor + 2$ intervals I's of the type $[a,b)$ such that the first $\lfloor T\rfloor + 1$ intervals are of length 1 and the last one is of length strictly smaller than 1 (eventually it is just a singleton). Then, any interval of the type $[t-1,t]$ for t in $[0,T]$ is included into the union of two such intervals. Therefore the event where all the N_I's are smaller than $u = \mathcal{N}/2$ is included into $\Omega_\mathcal{N}$. It remains to control the probability of the complementary of this event. By stationarity, all the first N_I's have the same distribution and satisfy Proposition 2. The last one can also be viewed as the truncation of a stationary point process to an interval of length smaller than 1. Therefore the exponential inequality of Proposition 2 also applies to the last interval. It remains to apply $\lfloor T \rfloor + 2$ times this exponential inequality and to use a union bound.

7.5.3. Proof of Corollary 4

As in the proof of Proposition 3, we use the notation $\square$. The non-asymptotic part of the result is just a pure application of Theorem 2, with the choices of B_φ and V_φ given by (5.5) and (5.6). The next step consists in controlling the martingale $\psi(\varphi)^2 \bullet (N-\Lambda)_T$ on $\Omega_{V,B}$. To do so, let us apply (7.7) to H such that for any m,

$$H_t^{(m)} = \psi_t^{(m)}(\varphi)^2 \mathbb{1}_{t \le \tau'},$$

with $B = B_\varphi^2$ and $\tau = T$ and where τ' is defined in (7.1) (see the proof of Theorem 2). The assumption to be fulfilled is checked as in the proof of Theorem 2. But as previously, on $\Omega_{V,B}$, $H \bullet (N-\Lambda)_T = \psi(\varphi)^2 \bullet (N-\Lambda)_T$ and also $H^2 \bullet \Lambda_T = \psi(\varphi)^4 \bullet \Lambda_T$. Moreover on $\Omega_{\mathcal{N}} \subset \Omega_{V,B}$

$$H^2 \bullet \Lambda_T = \psi(\varphi)^4 \bullet \Lambda_T \le v := TM(\max_m \nu^{(m)} + \mathcal{N} \max_{m,\ell} h_\ell^{(m)}) B_\varphi^4.$$

Recall that $x = \alpha \log(T)$. So on $\Omega_{V,B}$, with probability larger than $1-(M+KM^2)e^{-x} = 1-(M+KM^2)T^{-\alpha}$, one has that for all $\varphi \in \Phi$,

$$\psi(\varphi)^2 \bullet N_T \le \psi(\varphi)^2 \bullet \Lambda_T + \sqrt{2vx} + \frac{B_\varphi^2 x}{3}.$$

So that for all $\varphi \in \Phi$,

$$\psi(\varphi)^2 \bullet N_T \le \square_{M,f_0} \left[\mathcal{N} \|\varphi\|_T^2 + \|\Phi\|_\infty^2 \mathcal{N}^2 \sqrt{T\mathcal{N} \log(T)} \right].$$

Also, since $\mathcal{N} = \log^2(T)$, one can apply Corollary 3, with $\beta = \alpha$. We finally choose c as in Proposition 5. This leads to the result.

Acknowledgements: We are very grateful to Christine Tuleau-Malot who allowed us to use her R programs simulating Hawkes processes. The research of Patricia Reynaud-Bouret and Vincent Rivoirard is partly supported by the french Agence Nationale de la Recherche (ANR 2011 BS01 010 01 projet Calibration). The authors would like to thank the anonymous Associate Editor and Referees for helpful comments and suggestions.

References

[1] Aalen, O. (1980). *A model for nonparametric regression analysis of counting processes.* Mathematical statistics and probability theory (Proc. Sixth Internat. Conf., Wisła, 1978). *Lecture Notes in Statistics*, **2**, 1–25.

[2] Andersen, P. K., Borgan, Ø., Gill, R. D. and Keiding, N. (1993). *Statistical models based on counting processes. Springer Series in Statistics.* Springer-Verlag, New York.

[3] Bacry, E., Delattre, S., Hoffmann, M. and Muzy, J.F. (2013). Some limit theorems for Hawkes processes and application to financial statistics. *Stoch. Proc. Appl.* **123**(7), 2475–2499.

[4] Bercu, B. and Touati, A. (2008). Exponential inequalities for self-normalized martingales with applications. *Ann. Appl. Prob.* **18**(5), 1848–1869.

[5] Bertin, K., Le Pennec, E. and Rivoirard, V. (2011). Adaptive Dantzig density estimation. *Ann. Inst. Henri Poincaré Probab. Statist.* **47**(1), 43–74.

[6] Bickel, P. J., Ritov, Y. and Tsybakov, A. B. (2009). Simultaneous analysis of Lasso and Dantzig selector. *Ann. Statist.* **37**(4), 1705–1732.

[7] Bowsher, C. G. (2010). Stochastic kinetic models: Dynamic independence, modularity and graphs. *Ann. Statist.* **38**(4):2242–2281.

[8] Brémaud, P. (1981). *Point processes and queues.* Springer-Verlag, New York.

[9] Brémaud, P. and Massoulié, L. (1996). Stability of nonlinear Hawkes processes. *Ann. Prob.* **24**(3), 1563–1588.

[10] Brette, R. and Destexhe, A. (2012). *Handbook of Neural Activity Measurement.* Cambridge University Press.

[11] Brunel, E. and Comte, F. (2005). Penalized contrast estimation of density and hazard rate with censored data. *Sankhyā.* **67**(3), 441–475.

[12] Brunel, E. and Comte, F. (2008). Adaptive estimation of hazard rate with censored data. *Comm. Statist. Theory M.* **37**(8-10), 1284–1305.

[13] Bühlmann, P. and van de Geer, S. (2011). *Statistics for high-dimensional data.* Springer, Heidelberg.

[14] Bunea, F. and McKeague, I. W. (2005). Covariate selection for semiparametric hazard function regression models. *J. Multivariate Anal.* **92**, 186–204.

[15] Bunea, F., Tsybakov, A. B. and Wegkamp, M. H. (2006). Aggregation and sparsity via ℓ_1 penalized least squares. *Proc. 19th Annual Conf. Learning Theory (COLT 2006). Lecture Notes in Artificial Intelligence v.4005* (Lugosi, G. and Simon, H.U., eds.). Springer-Verlag, Berlin-Heidelberg.

[16] Bunea, F., Tsybakov, A. B. and Wegkamp, M. H. (2007). Sparse density estimation with l_1 penalties. *Lecture Notes in Artificial Intelligence.* **4539**, 530–543.

[17] Bunea, F., Tsybakov, A. B. and Wegkamp, M. H. (2007). Aggregation for Gaussian regression. *Ann. Statist.* **35**(4), 1674–1697.

[18] Bunea, F., Tsybakov, A. B. and Wegkamp, M. H. (2007). Sparsity Oracle Inequalities for the Lasso. *Elec. J. Statist.* **1**, 169–194.

[19] Candes, E. and Tao, T. (2007). The Dantzig selector: statistical estimation when p is much larger than n. *Ann. Statist.* **35**(6), 2313–2351.

[20] Carstensen, L., Sandelin, A., Winther, O. and Hansen, N. R. (2010). Multivariate Hawkes process models of the occurrence of regulatory elements and an analysis of the pilot ENCODE regions. *BMC Bioinformatics* **11**(456).

[21] Chagny, G. (2012). Adaptive warped kernel estimators. http://hal.

archives-ouvertes.fr/hal-00715184.

[22] Chornoboy, E. S., Schramm, L. P. and Karr, A. F. (1988). Maximum likelihood identification of neural point process systems. *Biol. Cybern.* **59**, 265–275.

[23] Comte, F., Gaïffas, S. and Guilloux A. (2011) Adaptive estimation of the conditional intensity of marker-dependent counting processes. *Ann. Inst. H. Poincaré Probab. Statist.* **47**(4), 1171-1196.

[24] Daley, D. J. and Vere-Jones, D. (2003). *An introduction to the theory of point processes. Vol. I.* Probability and its Applications. Second edition. Springer-Verlag, New York.

[25] de la Peña, V. H. (1999). A general class of exponential inequalities for martingales and ratios. *Ann. Prob.* **27**(1), 537–564.

[26] de la Peña, V. H., Lai, T. L., Shao, Q. M. (2009). *Self-Normalized Processes.* Springer-Verlag, Berlin Heidelberg.

[27] Dzhaparidze, K. and van Zanten, J. H. (2001). On Bernstein-type inequalities for martingales. *Stoch. Proc. Appl.* **93**(1), 109–117.

[28] Fu, W. J. (1998). Penalized regressions: the bridge versus the Lasso. *J. Comput. Graph. Statist.* **7**(3), 397–416.

[29] Gaïffas, S. and Guilloux A. (2012). High-dimensional additive hazard models and the Lasso. *Electron. J. Statist.* **6**, 522–546.

[30] Grégoire, G. (1993). Least squares cross-validation for counting process intensities. *Scand. J. Statist.* **20**(4), 343–360.

[31] Grün, S., Diesmann, M., Grammont, F., Riehle, A. and Aertsen, A. (1999). Detecting unitary events without discretization in time. *J. Neurosci. meth.* **94**(1), 67-79.

[32] Gusto, G. and Schbath, S. (2005). FADO: a statistical method to detect favored or avoided distances between occurrences of motifs using the Hawkes' model. *Stat. Appl. Gen. Mol.* **4**.

[33] Härdle, W., Kerkyacharian, G., Picard, D. and Tsybakov A. B. (1998). *Wavelets, Approximation and Statistical Applications.* Springer-Verlag, Berlin.

[34] Hawkes, A. G. (1971). Point spectra of some mutually exciting point processes. *J. R. Stat. Soc. Ser. B* **33**, 438–443.

[35] Huang, J., Ma, S. and Zhang, C. H. (2008). Adaptive Lasso for sparse high-dimensional regression models. *Statist. Sinica.* **18**, 1603–1618.

[36] Jacobsen, Martin (2006). *Point process theory and applications. Marked point and piecewise deterministic processes.* Probability and its Applications. Birkhäuser Boston, Inc., Boston, MA.

[37] Koltchinskii, V., Lounici, K. and Tsybakov, A.B. (2011). Nuclear norm penalization and optimal rates for noisy low rank matrix completion. http://lanl.arxiv.org/abs/1011.6256

[38] Krumin, M., Reutsky, I., and Shoham, S. (2010). Correlation-based analysis and generation of multiple spike trains using Hawkes models with an exogenous input. *Front. Comp. Neurosci.* **4**, article 147.

[39] Letue, F. (2000) *Modèle de Cox : Estimation par sélection de modèle et modèle de chocs bivarié*. PhD thesis.
[40] Liptser, R. and Spokoiny, V. (2000). Deviation probability bound for martingales with applications to statistical estimation. *Statist. Probabil. Lett.* **46**, 347–357.
[41] Massart, P. (2007). *Concentration inequalities and model selection.* Lectures from the 33rd Summer School on Probability Theory held in Saint-Flour, July 6–23, 2003. Springer, Berlin.
[42] Meinshausen, N. (2007). Relaxed Lasso. *Comput. Statist. Data An.* **52**(1), 374–393.
[43] Masud, M. S. and Borisyuk, R. (2011). Statistical technique for analysing functional connectivity of multiple spike trains. *J. Neurosci. Meth.* **196**(1), 201–219.
[44] Mitchell, L. and Cates, M.E. (2010). Hawkes process as a model of social interactions: a view on video dynamics. *J. Phys.* **A43**(4), 045101.
[45] Pernice, V., Staude, B., Cardanobile, S., and Rotter, S. (2011). How structure determines correlations in neuronal networks. *PLoS Comput. Biol.* **7**(5), e1002059.
[46] Pernice, V., Staude, B., Cardanobile, S., and Rotter, S. (2012). Recurrent interactions in spiking networks with arbitrary topology. *Phys. rev. E* 85:031916.
[47] Pillow, J. W., Shlens, J., Paninski, L., Sher, A., Litke, A. M., Chichilnisky, E. J. and Simoncelli, E. P. (2008). Spatio-temporal correlations and visual signalling in a complete neuronal population. *Nature* **454**, 995–999.
[48] Reynaud-Bouret, P. (2003). Adaptive estimation of the intensity of inhomogeneous Poisson processes via concentration inequalities. *Probab. Theory Rel.* **126**(1), 103–153.
[49] Reynaud-Bouret, P. (2006). Penalized projection estimators of the Aalen multiplicative intensity. *Bernoulli* **12**(4), 633–661.
[50] Reynaud-Bouret, P. and Rivoirard, V. (2010). Near optimal thresholding estimation of a Poisson intensity on the real line. *Elec. J. Statist.* **4**, 172–238.
[51] Reynaud-Bouret, P. and Roy E. (2007). Some non asymptotic tail estimates for Hawkes processes. *Bull. Belg. Math. Soc.* **13**(5), 883–896.
[52] Reynaud-Bouret, P. and Schbath, S. (2010). Adaptive estimation for Hawkes processes; application to genome analysis. *Ann. Statist.* **38**(5), 2781–2822.
[53] Reynaud-Bouret, P., Tuleau-Malot, C., Rivoirard, V. and Grammont, F. (2013). Spike trains as (in)homogeneous Poisson processes or Hawkes processes: nonparametric adaptive estimation and goodness-of-fit tests. http://hal.archives-ouvertes.fr/hal-00789127
[54] Rudelson, M. and Vershynin, R. (2008). On sparse reconstruction from Fourier and Gaussian measurements. *Comm. Pure Appl. Math.* **61**(8), 1025–1045.
[55] Rudelson, M. and Vershynin, R. (2009). Smallest singular value of a random rectangular matrix. *Comm. Pure Appl. Math.* **62**(12), 1707–1739.

[56] Rudelson, M. and Vershynin, R. (2010). Non-asymptotic theory of random matrices: extreme singular values. *Proc. Int. Congress Math.* Volume III, Hindustan Book Agency, New Delhi, 1576–1602.

[57] Shorack, G. R. and Wellner, J. A. (1986) *Empirical processes with applications to statistics.* Wiley Series in Probability and Mathematical Statistics: Probability and Mathematical Statistics. John Wiley & Sons Inc., New York.

[58] Tibshirani, R. (1996). Regression shrinkage and selection via the Lasso. *J. R. Stat. Soc. Ser. B* **58**(1), 267–288.

[59] van de Geer, S. (1995). Exponential inequalities for martingales, with application to maximum likelihood estimation for counting processes. *Ann. Statist.* **23**(5), 1779–1801.

[60] van de Geer, S. (2008). High dimensional generalized linear models and the Lasso. *Ann. Statist.* **36**(2), 614–645.

[61] van de Geer, S. and Bühlmann, P. and Zhou, S. (2011). The adaptive and the thresholded Lasso for potentially misspecified models (and a lower bound for the Lasso). *Elec. J. Statist.* **5**, 688-749.

[62] Vere-Jones, D. and Ozaki, T. (1982). Some examples of statistical estimation applied to earthquake data I: cyclic Poisson and self-exciting models. *Ann. I. Stat. Math.* **34**(1), 189–207.

[63] Willett, R.M. and Nowak, R.D. (2007). Multiscale Poisson Intensity and Density Estimation. *IEEE T. Inform. Theory* **53**(9), 3171–3187.

[64] Zou, H. (2006). The adaptive Lasso and its oracle properties. *J. Am. Statist. Assoc.* **101**(476), 1418–1429.